# manipulating-data

## Jian

### 16/07/2020

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.2     v dplyr   1.0.0
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

load data

```r
surveys <- read_csv('data_raw/portal_data_joined.csv')
```

```
## Parsed with column specification:
## cols(
##   record_id = col_double(),
##   month = col_double(),
##   day = col_double(),
##   year = col_double(),
##   plot_id = col_double(),
##   species_id = col_character(),
##   sex = col_character(),
##   hindfoot_length = col_double(),
##   weight = col_double(),
##   genus = col_character(),
##   species = col_character(),
##   taxa = col_character(),
##   plot_type = col_character()
## )
```

```r
str(surveys)
```

```
## tibble [34,786 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ record_id      : num [1:34786] 1 72 224 266 349 363 435 506 588 661 ...
##  $ month          : num [1:34786] 7 8 9 10 11 11 12 1 2 3 ...
```

```
##  $ day           : num [1:34786] 16 19 13 16 12 12 10 8 18 11 ...
##  $ year          : num [1:34786] 1977 1977 1977 1977 1977 ...
##  $ plot_id       : num [1:34786] 2 2 2 2 2 2 2 2 2 2 ...
##  $ species_id    : chr [1:34786] "NL" "NL" "NL" "NL" ...
##  $ sex           : chr [1:34786] "M" "M" NA NA ...
##  $ hindfoot_length: num [1:34786] 32 31 NA NA NA NA NA NA NA NA ...
##  $ weight        : num [1:34786] NA NA NA NA NA NA NA NA 218 NA ...
##  $ genus         : chr [1:34786] "Neotoma" "Neotoma" "Neotoma" "Neotoma" ...
##  $ species       : chr [1:34786] "albigula" "albigula" "albigula" "albigula" ...
##  $ taxa          : chr [1:34786] "Rodent" "Rodent" "Rodent" "Rodent" ...
##  $ plot_type     : chr [1:34786] "Control" "Control" "Control" "Control" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   record_id = col_double(),
##   ..   month = col_double(),
##   ..   day = col_double(),
##   ..   year = col_double(),
##   ..   plot_id = col_double(),
##   ..   species_id = col_character(),
##   ..   sex = col_character(),
##   ..   hindfoot_length = col_double(),
##   ..   weight = col_double(),
##   ..   genus = col_character(),
##   ..   species = col_character(),
##   ..   taxa = col_character(),
##   ..   plot_type = col_character()
##   .. )
```

```r
View(surveys)
```

subset columns

```r
select(surveys, 'plot_id', 'species_id', 'weight')
```

```
## Warning: '...' is not empty.
##
## We detected these problematic arguments:
## * 'needs_dots'
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 34,786 x 3
##    plot_id species_id weight
##      <dbl> <chr>       <dbl>
## 1        2 NL             NA
## 2        2 NL             NA
## 3        2 NL             NA
## 4        2 NL             NA
## 5        2 NL             NA
## 6        2 NL             NA
## 7        2 NL             NA
## 8        2 NL             NA
```

```
##  9        2 NL           218
## 10        2 NL            NA
## # ... with 34,776 more rows
```

or

```r
select(surveys, -c('record_id', 'species_id'))
```

```
## Warning: '...' is not empty.
##
## We detected these problematic arguments:
## * 'needs_dots'
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 34,786 x 11
##    month   day  year plot_id sex   hindfoot_length weight genus species taxa
##    <dbl> <dbl> <dbl>   <dbl> <chr>           <dbl>  <dbl> <chr> <chr>   <chr>
## 1      7    16  1977       2 M                  32     NA Neot~ albigu~ Rode~
## 2      8    19  1977       2 M                  31     NA Neot~ albigu~ Rode~
## 3      9    13  1977       2 <NA>               NA     NA Neot~ albigu~ Rode~
## 4     10    16  1977       2 <NA>               NA     NA Neot~ albigu~ Rode~
## 5     11    12  1977       2 <NA>               NA     NA Neot~ albigu~ Rode~
## 6     11    12  1977       2 <NA>               NA     NA Neot~ albigu~ Rode~
## 7     12    10  1977       2 <NA>               NA     NA Neot~ albigu~ Rode~
## 8      1     8  1978       2 <NA>               NA     NA Neot~ albigu~ Rode~
## 9      2    18  1978       2 M                  NA    218 Neot~ albigu~ Rode~
## 10     3    11  1978       2 <NA>               NA     NA Neot~ albigu~ Rode~
## # ... with 34,776 more rows, and 1 more variable: plot_type <chr>
```

choose rows on specific conditions

```r
filter(surveys, year == 1995)
```

```
## Warning: '...' is not empty.
##
## We detected these problematic arguments:
## * 'needs_dots'
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 1,180 x 13
##    record_id month   day  year plot_id species_id sex   hindfoot_length weight
##        <dbl> <dbl> <dbl> <dbl>   <dbl> <chr>      <chr>           <dbl>  <dbl>
## 1      22314     6     7  1995       2 NL         M                  34     NA
## 2      22728     9    23  1995       2 NL         F                  32    165
## 3      22899    10    28  1995       2 NL         F                  32    171
## 4      23032    12     2  1995       2 NL         F                  33     NA
## 5      22003     1    11  1995       2 DM         M                  37     41
## 6      22042     2     4  1995       2 DM         F                  36     45
```

```
## 7       22044    2    4  1995     2 DM        M                   37    46
## 8       22105    3    4  1995     2 DM        F                   37    49
## 9       22109    3    4  1995     2 DM        M                   37    46
## 10      22168    4    1  1995     2 DM        M                   36    48
## # ... with 1,170 more rows, and 4 more variables: genus <chr>, species <chr>,
## #   taxa <chr>, plot_type <chr>
```

```r
surveys2 <- filter(surveys, weight < 5)
```

```r
surveySml <- select(surveys2, 'species_id', 'sex', 'weight')
```

```r
surveySml <- select(filter(surveys, weight < 5), 'species_id', 'sex', 'weight')
```

pipes

```r
surveySml <- surveys %>%
  filter(weight < 5) %>%
  select('species_id', 'sex', 'weight')
```

```r
surveySml
```

```
## Warning: '...' is not empty.
##
## We detected these problematic arguments:
## * 'needs_dots'
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 17 x 3
##    species_id sex    weight
##    <chr>      <chr>  <dbl>
##  1 PF         F         4
##  2 PF         F         4
##  3 PF         M         4
##  4 RM         F         4
##  5 RM         M         4
##  6 PF         <NA>      4
##  7 PP         M         4
##  8 RM         M         4
##  9 RM         M         4
## 10 RM         M         4
## 11 PF         M         4
## 12 PF         F         4
## 13 RM         M         4
## 14 RM         M         4
## 15 RM         F         4
## 16 RM         M         4
## 17 RM         M         4
```

```r
surveys %>%
  filter(year < 1995) %>%
  select(year, sex, weight)
```

```
## Warning: '...' is not empty.
##
## We detected these problematic arguments:
## * 'needs_dots'
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 21,486 x 3
##     year sex    weight
##    <dbl> <chr>   <dbl>
##  1  1977 M          NA
##  2  1977 M          NA
##  3  1977 <NA>       NA
##  4  1977 <NA>       NA
##  5  1977 <NA>       NA
##  6  1977 <NA>       NA
##  7  1977 <NA>       NA
##  8  1978 <NA>       NA
##  9  1978 M         218
## 10  1978 <NA>       NA
## # ... with 21,476 more rows
```

create new columns

```r
surveys %>%
  mutate(weight_kg = weight / 1000) %>%
  head()
```

```
## Warning: '...' is not empty.
##
## We detected these problematic arguments:
## * 'needs_dots'
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 6 x 14
##   record_id month   day  year plot_id species_id sex   hindfoot_length weight
##       <dbl> <dbl> <dbl> <dbl>   <dbl> <chr>      <chr>            <dbl>  <dbl>
## 1         1     7    16  1977       2 NL         M                   32     NA
## 2        72     8    19  1977       2 NL         M                   31     NA
## 3       224     9    13  1977       2 NL         <NA>                NA     NA
## 4       266    10    16  1977       2 NL         <NA>                NA     NA
## 5       349    11    12  1977       2 NL         <NA>                NA     NA
## 6       363    11    12  1977       2 NL         <NA>                NA     NA
## # ... with 5 more variables: genus <chr>, species <chr>, taxa <chr>,
## #   plot_type <chr>, weight_kg <dbl>
```

```
surveys %>%
  mutate(weight_kg = weight / 1000,
         weight_lb = weight_kg * 2.2) %>%
  head()
```

```
## Warning: '...' is not empty.
##
## We detected these problematic arguments:
## * 'needs_dots'
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 6 x 15
##   record_id month   day  year plot_id species_id sex    hindfoot_length weight
##       <dbl> <dbl> <dbl> <dbl>   <dbl> <chr>      <chr>            <dbl>  <dbl>
## 1         1     7    16  1977       2 NL         M                   32     NA
## 2        72     8    19  1977       2 NL         M                   31     NA
## 3       224     9    13  1977       2 NL         <NA>                NA     NA
## 4       266    10    16  1977       2 NL         <NA>                NA     NA
## 5       349    11    12  1977       2 NL         <NA>                NA     NA
## 6       363    11    12  1977       2 NL         <NA>                NA     NA
## # ... with 6 more variables: genus <chr>, species <chr>, taxa <chr>,
## #   plot_type <chr>, weight_kg <dbl>, weight_lb <dbl>
```

```
surveys %>%
  filter(!is.na(weight)) %>%
  mutate(weight_kg = weight / 1000) %>%
  head()
```

```
## Warning: '...' is not empty.
##
## We detected these problematic arguments:
## * 'needs_dots'
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 6 x 14
##   record_id month   day  year plot_id species_id sex    hindfoot_length weight
##       <dbl> <dbl> <dbl> <dbl>   <dbl> <chr>      <chr>            <dbl>  <dbl>
## 1       588     2    18  1978       2 NL         M                   NA    218
## 2       845     5     6  1978       2 NL         M                   32    204
## 3       990     6     9  1978       2 NL         M                   NA    200
## 4      1164     8     5  1978       2 NL         M                   34    199
## 5      1261     9     4  1978       2 NL         M                   32    197
## 6      1453    11     5  1978       2 NL         M                   NA    218
## # ... with 5 more variables: genus <chr>, species <chr>, taxa <chr>,
## #   plot_type <chr>, weight_kg <dbl>
```

```
surveys %>%
  filter(!is.na(hindfoot_length)) %>%
  mutate(hindfoot_cm = hindfoot_length / 10) %>%
  filter(hindfoot_cm < 3) %>%
  select(species_id, hindfoot_cm)
```

```
## Warning: '...' is not empty.
##
## We detected these problematic arguments:
## * 'needs_dots'
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 15,371 x 2
##    species_id hindfoot_cm
##    <chr>           <dbl>
##  1 NL                2.8
##  2 NL                2.1
##  3 NL                2.1
##  4 NL                2.9
##  5 NL                2.9
##  6 DM                2.5
##  7 PF                1.5
##  8 PF                1.6
##  9 PF                1.5
## 10 PF                1.4
## # ... with 15,361 more rows
```

calculate the mean weight for each of the categories in sex

```
surveys %>%
  group_by(sex) %>%
  summarise(meanWeight = mean(weight, na.rm = TRUE))
```

```
## 'summarise()' ungrouping output (override with '.groups' argument)
```

```
## Warning: '...' is not empty.
##
## We detected these problematic arguments:
## * 'needs_dots'
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 3 x 2
##   sex   meanWeight
##   <chr>      <dbl>
## 1 F           42.2
## 2 M           43.0
## 3 <NA>        64.7
```

```
surveys %>%
  group_by(sex, species_id) %>%
  summarise(meanWeight = mean(weight, na.rm = TRUE)) %>%
  tail()
```

## `summarise()` regrouping output by 'sex' (override with `.groups` argument)

## Warning: `...` is not empty.
##
## We detected these problematic arguments:
## * `needs_dots`
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?

## # A tibble: 6 x 3
## # Groups:   sex [1]
##    sex   species_id meanWeight
##    <chr> <chr>           <dbl>
## 1 <NA>  SU                NaN
## 2 <NA>  UL                NaN
## 3 <NA>  UP                NaN
## 4 <NA>  UR                NaN
## 5 <NA>  US                NaN
## 6 <NA>  ZL                NaN

```
surveys %>%
  filter(!is.na(weight)) %>%
  group_by(sex, species_id) %>%
  summarise(meanWeight = mean(weight)) %>%
  tail()
```

## `summarise()` regrouping output by 'sex' (override with `.groups` argument)

## Warning: `...` is not empty.
##
## We detected these problematic arguments:
## * `needs_dots`
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?

## # A tibble: 6 x 3
## # Groups:   sex [1]
##    sex   species_id meanWeight
##    <chr> <chr>           <dbl>
## 1 <NA>  PL               25
## 2 <NA>  PM               20.2
## 3 <NA>  PP               14.6
## 4 <NA>  RM               11.1
## 5 <NA>  SF               40.5
## 6 <NA>  SH              130

```

```
surveys %>%
  filter(!is.na(weight)) %>%
  group_by(sex, species_id) %>%
  summarise(maxWeight = max(weight),
            minWeight = min(weight))
```

## 'summarise()' regrouping output by 'sex' (override with '.groups' argument)

## Warning: '...' is not empty.
##
## We detected these problematic arguments:
## * 'needs_dots'
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?

## # A tibble: 64 x 4
## # Groups:   sex [3]
##    sex   species_id maxWeight minWeight
##    <chr> <chr>          <dbl>     <dbl>
##  1 F     BA                18         6
##  2 F     DM                66        10
##  3 F     DO                76        12
##  4 F     DS               190        45
##  5 F     NL               274        32
##  6 F     OL                56        10
##  7 F     OT                46         5
##  8 F     OX                22        20
##  9 F     PB                54        12
## 10 F     PE                38        11
## # ... with 54 more rows
```

```
surveys %>%
  filter(!is.na(weight)) %>%
  group_by(sex, species_id) %>%
  summarise(maxWeight = max(weight),
            minWeight = min(weight)) %>%
  arrange(minWeight)
```

## 'summarise()' regrouping output by 'sex' (override with '.groups' argument)

## Warning: '...' is not empty.
##
## We detected these problematic arguments:
## * 'needs_dots'
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?

## # A tibble: 64 x 4
## # Groups:   sex [3]

```
##    sex   species_id maxWeight minWeight
##    <chr> <chr>          <dbl>     <dbl>
##  1 F     PF                24         4
##  2 F     RM                21         4
##  3 M     PF                25         4
##  4 M     PP                42         4
##  5 M     RM                29         4
##  6 <NA>  PF                 8         4
##  7 F     OT                46         5
##  8 F     PP                74         5
##  9 F     BA                18         6
## 10 M     BA                 9         6
## # ... with 54 more rows
```

```r
surveys %>%
  filter(!is.na(weight)) %>%
  group_by(sex, species_id) %>%
  summarise(meanWeight = mean(weight),
            maxWeight = max(weight),
            minWeight = min(weight)) %>%
  arrange(desc(meanWeight))
```

```
## 'summarise()' regrouping output by 'sex' (override with '.groups' argument)
```

```
## Warning: '...' is not empty.
##
## We detected these problematic arguments:
## * 'needs_dots'
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 64 x 5
## # Groups:   sex [3]
##    sex   species_id meanWeight maxWeight minWeight
##    <chr> <chr>           <dbl>     <dbl>     <dbl>
##  1 <NA>  NL              168.        243        83
##  2 M     NL              166.        280        30
##  3 F     NL              154.        274        32
##  4 M     SS              130         130       130
##  5 <NA>  SH              130         130       130
##  6 M     DS              122.        170        12
##  7 <NA>  DS              120         152        78
##  8 F     DS              118.        190        45
##  9 F     SH               78.8       140        30
## 10 F     SF               69         199        46
## # ... with 54 more rows
```

```r
surveys %>%
  count(sex)
```

```
## Warning: '...' is not empty.
```

```
##
## We detected these problematic arguments:
## * `needs_dots`
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?

## # A tibble: 3 x 2
##   sex       n
##   <chr> <int>
## 1 F     15690
## 2 M     17348
## 3 <NA>   1748
```

```r
surveys %>%
  group_by(sex) %>%
  summarise(count = n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)

## Warning: `...` is not empty.
##
## We detected these problematic arguments:
## * `needs_dots`
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?

## # A tibble: 3 x 2
##   sex   count
##   <chr> <int>
## 1 F     15690
## 2 M     17348
## 3 <NA>   1748
```

```r
surveys %>%
  count(sex, sort = TRUE)
```

```
## Warning: `...` is not empty.
##
## We detected these problematic arguments:
## * `needs_dots`
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?

## # A tibble: 3 x 2
##   sex       n
##   <chr> <int>
## 1 M     17348
## 2 F     15690
## 3 <NA>   1748
```

```
surveys %>%
  count(sex, species) %>%
  View()
```

```
surveys %>%
  count(sex, species) %>%
  arrange(species, desc(n))
```

```
## Warning: '...' is not empty.
##
## We detected these problematic arguments:
## * 'needs_dots'
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 81 x 3
##    sex   species            n
##    <chr> <chr>          <int>
##  1 F     albigula         675
##  2 M     albigula         502
##  3 <NA>  albigula          75
##  4 <NA>  audubonii         75
##  5 F     baileyi         1646
##  6 M     baileyi         1216
##  7 <NA>  baileyi           29
##  8 <NA>  bilineata        303
##  9 <NA>  brunneicapillus   50
## 10 <NA>  chlorurus         39
## # ... with 71 more rows
```

```
surveys %>%
  count(plot_type)
```

```
## Warning: '...' is not empty.
##
## We detected these problematic arguments:
## * 'needs_dots'
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 5 x 2
##   plot_type                    n
##   <chr>                    <int>
## 1 Control                  15611
## 2 Long-term Krat Exclosure  5118
## 3 Rodent Exclosure          4233
## 4 Short-term Krat Exclosure  5906
## 5 Spectab exclosure         3918
```

```
surveys %>%
  filter(!is.na(hindfoot_length)) %>%
  group_by(species_id) %>%
  summarise(meanHindLen = mean(hindfoot_length),
            minHindLen = min(hindfoot_length),
            maxHindLen = max(hindfoot_length),
            Count = n()) %>%
  arrange(species_id, Count) # how can I arrange count in descending order?
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## Warning: `...` is not empty.
##
## We detected these problematic arguments:
## * `needs_dots`
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 25 x 5
##    species_id meanHindLen minHindLen maxHindLen Count
##    <chr>            <dbl>      <dbl>      <dbl> <int>
##  1 AH                  33         31         35     2
##  2 BA                  13          6         16    45
##  3 DM                36.0         16         50  9972
##  4 DO                35.6         26         64  2887
##  5 DS                49.9         39         58  2132
##  6 NL                32.3         21         70  1074
##  7 OL                20.5         12         39   920
##  8 OT                20.3         13         50  2139
##  9 OX                19.1         13         21     8
## 10 PB                26.1          2         47  2864
## # ... with 15 more rows
```

```
surveys %>%
  filter(!is.na(weight)) %>%
  group_by(year) %>%
  filter(weight == max(weight)) %>%
  select(year, genus, species_id, weight) %>%
  arrange(year)
```

```
## Warning: `...` is not empty.
##
## We detected these problematic arguments:
## * `needs_dots`
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 27 x 4
## # Groups:   year [26]
```

```
##      year genus     species_id weight
##     <dbl> <chr>     <chr>       <dbl>
##  1  1977 Dipodomys DS            149
##  2  1978 Neotoma   NL            232
##  3  1978 Neotoma   NL            232
##  4  1979 Neotoma   NL            274
##  5  1980 Neotoma   NL            243
##  6  1981 Neotoma   NL            264
##  7  1982 Neotoma   NL            252
##  8  1983 Neotoma   NL            256
##  9  1984 Neotoma   NL            259
## 10  1985 Neotoma   NL            225
## # ... with 17 more rows
```

```r
surveysGenusWeight <- surveys %>%
  filter(!is.na(weight)) %>%
  group_by(plot_id, genus) %>%
  summarise(meanWeight = mean(weight))
```

```
## `summarise()` regrouping output by 'plot_id' (override with `.groups` argument)
```

```r
view(surveysGenusWeight)
```

```r
surveys_spread <- surveysGenusWeight %>%
  spread(key = genus, value = meanWeight)
```

```r
surveys_spread <- surveysGenusWeight %>%
  spread(genus, meanWeight, fill = 0)
view(surveys_spread)
```

```r
surveys_gather <- surveys_spread %>%
  gather(key = species_id, value = meanWeight, -plot_id)
view(surveys_gather)
```

```r
surveys_gather <- surveys_spread %>%
  gather(key = species_id, value = meanWeight, Baiomys:Spermophilus)
view(surveys_gather)
```

```r
genera_spread <- surveys %>%
  group_by(plot_id, year) %>%
  summarise(n_genera = n_distinct(genus)) %>%
  spread(key = year, value = n_genera)
```

```
## `summarise()` regrouping output by 'plot_id' (override with `.groups` argument)
```

```r
View(genera_spread)
```

```r
genera_gather <- genera_spread %>%
  gather(key = year, value = n_genera, -plot_id)

View(genera_gather)
```

14

```
surveys_measurement <- surveys %>%
  gather(key = 'measurement', value = 'value', hindfoot_length, weight)
View(surveys_measurement)
```

```
surveys_mean_measure <-  surveys_measurement %>%
  group_by(year, measurement, plot_type) %>%
  summarise(mean_value = mean(value, na.rm = TRUE)) %>%
  spread(measurement, mean_value)
```

```
## `summarise()` regrouping output by 'year', 'measurement' (override with `.groups` argument)
```

```
View(surveys_mean_measure)
```

## Export data

**remove invalid data**

```
surveys_complete <- surveys %>%
  filter(!is.na(hindfoot_length),
         !is.na(weight),
         !is.na(sex))
```

**extract the most common species (n >= 50)**

```
species_count <- surveys_complete %>%
  count(species_id) %>%
  filter(n >= 50)
```

**only keep the most common species**

```
surveys_complete <- surveys_complete %>%
  filter(species_id %in% species_count$species_id)

dim(surveys_complete)
```

```
## [1] 30463     13
```

```
View(surveys_complete)
```

```
write_csv(surveys_complete, path = 'data/surveys_complete.csv')
```

The end