R workshop_4

Jackie Maud

21 January 2019

```
library(tinytex)
```

Creating objects in R

```
3+5
```

```
## [1] 8
```

```
#assign value to object

weight_kg<-55

# to convert to pounds

weight_lb <-2.2 * weight_kg


sqrt(weight_kg)
```

```
## [1] 7.416198
```

```
# rounds off to required number sigfig)

round(3.14159) #default is 3 sigfig
```

```
## [1] 3
```

```
round(3.14159, digits=2)
```

```
## [1] 3.14
```

```
#or
round(3.14159, 2)
```

```
## [1] 3.14
```

Vectors and data types

Some basic data types in R

```
weight_g<-c(50, 60, 65, 82)

animals<-c("mouse", "rat", "dog")
```

Different types of vector (i.e. atomic - only one type of data)

numeric character *logical (TRUE or FALSE)* factors (categorical) *dates

A vector is a data structure (has a linear structure)

Other data structures:

lists data frames *matrices

Data frames

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0
## -- Conflicts ------------------------------------------------------------------------------------ tidy
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#utils::  tells us which source to get package from

#Dowmload data onto computer, need next function to read the data

download.file(url="https://ndownloader.figshare.com/files/2292169", destfile = "read_data/portal_data_j

library(here) #This package makes working directories and file paths made easy
```

```
## here() starts at C:/Users/HP/Documents/R/R projects_Jackie/Jackie_zooplankton_diets
```

```
#

surveys<-read_csv(here("read_data", "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()
## )
```

```
surveys
```

```
## # A tibble: 34,786 x 13
##    record_id month   day  year plot_id species_id sex   hindfoot_length
##        <dbl> <dbl> <dbl> <dbl>   <dbl> <chr>      <chr>           <dbl>
##  1         1     7    16  1977       2 NL         M                  32
##  2        72     8    19  1977       2 NL         M                  31
##  3       224     9    13  1977       2 NL         <NA>               NA
```

```
##  4       266    10    16  1977       2 NL         <NA>               NA
##  5       349    11    12  1977       2 NL         <NA>               NA
##  6       363    11    12  1977       2 NL         <NA>               NA
##  7       435    12    10  1977       2 NL         <NA>               NA
##  8       506     1     8  1978       2 NL         <NA>               NA
##  9       588     2    18  1978       2 NL         M                  NA
## 10       661     3    11  1978       2 NL         <NA>               NA
## # ... with 34,776 more rows, and 5 more variables: weight <dbl>,
## #   genus <chr>, species <chr>, taxa <chr>, plot_type <chr>
```

```
#columns of data frame are vectors

#data frame as vectors pf equal length
#matrix can have columns of different length

str(surveys)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 34786 obs. of  13 variables:
##  $ record_id      : num  1 72 224 266 349 363 435 506 588 661 ...
##  $ month          : num  7 8 9 10 11 11 12 1 2 3 ...
##  $ day            : num  16 19 13 16 12 12 10 8 18 11 ...
##  $ year           : num  1977 1977 1977 1977 1977 ...
##  $ plot_id        : num  2 2 2 2 2 2 2 2 2 2 ...
##  $ species_id     : chr  "NL" "NL" "NL" "NL" ...
##  $ sex            : chr  "M" "M" NA NA ...
##  $ hindfoot_length: num  32 31 NA NA NA NA NA NA NA NA ...
##  $ weight         : num  NA NA NA NA NA NA NA NA 218 NA ...
##  $ genus          : chr  "Neotoma" "Neotoma" "Neotoma" "Neotoma" ...
##  $ species        : chr  "albigula" "albigula" "albigula" "albigula" ...
##  $ taxa           : chr  "Rodent" "Rodent" "Rodent" "Rodent" ...
##  $ plot_type      : chr  "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()
##   .. )
```

```
dim(surveys)
```

```
## [1] 34786    13
```

```
nrow(surveys)
```

```
## [1] 34786
```

```
summary(surveys)
```

```
##    record_id         month             day            year
##  Min.   :    1   Min.   : 1.000   Min.   : 1.0   Min.   :1977
##  1st Qu.: 8964   1st Qu.: 4.000   1st Qu.: 9.0   1st Qu.:1984
##  Median :17762   Median : 6.000   Median :16.0   Median :1990
##  Mean   :17804   Mean   : 6.474   Mean   :16.1   Mean   :1990
##  3rd Qu.:26655   3rd Qu.:10.000   3rd Qu.:23.0   3rd Qu.:1997
##  Max.   :35548   Max.   :12.000   Max.   :31.0   Max.   :2002
##
##     plot_id       species_id           sex            hindfoot_length
##  Min.   : 1.00   Length:34786       Length:34786       Min.   : 2.00
##  1st Qu.: 5.00   Class :character   Class :character   1st Qu.:21.00
##  Median :11.00   Mode  :character   Mode  :character   Median :32.00
##  Mean   :11.34                                         Mean   :29.29
##  3rd Qu.:17.00                                         3rd Qu.:36.00
##  Max.   :24.00                                         Max.   :70.00
##                                                        NA's   :3348
##      weight          genus              species              taxa
##  Min.   :  4.00   Length:34786       Length:34786       Length:34786
##  1st Qu.: 20.00   Class :character   Class :character   Class :character
##  Median : 37.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 42.67
##  3rd Qu.: 48.00
##  Max.   :280.00
##  NA's   :2503
##   plot_type
##  Length:34786
##  Class :character
##  Mode  :character
##
##
##
##
```

Indexing and subsetting data frames

First using square brackets []

Square brackets are great for defining

Do restart and reload (green down arrow)

```
#First define the row coordinate, and then the column, Also write row and then column

surveys[1,1]
```

```
## # A tibble: 1 x 1
##   record_id
##       <dbl>
## 1         1
```

```
surveys[1,6]
```

```
## # A tibble: 1 x 1
##   species_id
##   <chr>
```

```
## 1 NL
```

```
#Defining only which element we want will return a data frame

surveys[1]
```

```
## # A tibble: 34,786 x 1
##    record_id
##        <dbl>
##  1         1
##  2        72
##  3       224
##  4       266
##  5       349
##  6       363
##  7       435
##  8       506
##  9       588
## 10       661
## # ... with 34,776 more rows
```

```
surveys[1:3, 7]
```

```
## # A tibble: 3 x 1
##   sex
##   <chr>
## 1 M
## 2 M
## 3 <NA>
```

```
# Give us all rows and columns except column 7

surveys[, -c(1:5)]
```

```
## # A tibble: 34,786 x 8
##    species_id sex   hindfoot_length weight genus   species  taxa  plot_type
##    <chr>      <chr>           <dbl>  <dbl> <chr>   <chr>    <chr> <chr>
##  1 NL         M                  32     NA Neotoma albigula Rode~ Control
##  2 NL         M                  31     NA Neotoma albigula Rode~ Control
##  3 NL         <NA>               NA     NA Neotoma albigula Rode~ Control
##  4 NL         <NA>               NA     NA Neotoma albigula Rode~ Control
##  5 NL         <NA>               NA     NA Neotoma albigula Rode~ Control
##  6 NL         <NA>               NA     NA Neotoma albigula Rode~ Control
##  7 NL         <NA>               NA     NA Neotoma albigula Rode~ Control
##  8 NL         <NA>               NA     NA Neotoma albigula Rode~ Control
##  9 NL         M                  NA    218 Neotoma albigula Rode~ Control
## 10 NL         <NA>               NA     NA Neotoma albigula Rode~ Control
## # ... with 34,776 more rows
```

Data manipulation (tidyverse)

Key functions for data manipulation

select(): subsetting columns filter(): subsets of rows based on conditions *mutate(): create new columns, based on info from other columns* group_by(): creates groups based on categorical data *summarize(): create summary stats on grouped data* arrange(): sort results *count(): gives a count of discrete values

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

```
## # 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
```

```
select(surveys, -record_id)
```

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

```
#use == for logical statements

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

```
## # A tibble: 1,180 x 13
##    record_id month   day  year plot_id species_id sex   hindfoot_length
##        <dbl> <dbl> <dbl> <dbl>   <dbl> <chr>      <chr>           <dbl>
##  1     22314     6     7  1995       2 NL         M                  34
##  2     22728     9    23  1995       2 NL         F                  32
##  3     22899    10    28  1995       2 NL         F                  32
##  4     23032    12     2  1995       2 NL         F                  33
##  5     22003     1    11  1995       2 DM         M                  37
##  6     22042     2     4  1995       2 DM         F                  36
##  7     22044     2     4  1995       2 DM         M                  37
##  8     22105     3     4  1995       2 DM         F                  37
##  9     22109     3     4  1995       2 DM         M                  37
## 10     22168     4     1  1995       2 DM         M                  36
## # ... with 1,170 more rows, and 5 more variables: weight <dbl>,
## #   genus <chr>, species <chr>, taxa <chr>, plot_type <chr>
```

```
filter(surveys, year==1995,
       species_id=="NL")
```

```
## # A tibble: 8 x 13
##   record_id month   day  year plot_id species_id sex   hindfoot_length
##       <dbl> <dbl> <dbl> <dbl>   <dbl> <chr>      <chr>           <dbl>
## 1     22314     6     7  1995       2 NL         M                  34
## 2     22728     9    23  1995       2 NL         F                  32
## 3     22899    10    28  1995       2 NL         F                  32
## 4     23032    12     2  1995       2 NL         F                  33
## 5     22847    10    28  1995      12 NL         M                  34
## 6     22998    12     2  1995      12 NL         M                  33
## 7     23124    12    21  1995      12 NL         F                  32
## 8     22476     7    20  1995      24 NL         F                  31
## # ... with 5 more variables: weight <dbl>, genus <chr>, species <chr>,
## #   taxa <chr>, plot_type <chr>
```

Pipes

Pipe: ctrl-shift-m

Write multiple arguments in a sentence

```
surveys %>%
  filter(weight<5) %>%
  select(species_id, sex, weight)
```

```
## # 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
```

```
surveys_sml<-surveys %>%
filter(weight<5) %>%
  select(species_id, sex, weight)
```

Challenge 1

All data for 1995 and year, sex and weight

```
surveys_jlm<- surveys %>%
  filter(year==1995) %>%
  select(year, sex, weight)

surveys_jlm
```

```
## # A tibble: 1,180 x 3
##     year sex   weight
##    <dbl> <chr>  <dbl>
##  1  1995 M         NA
##  2  1995 F        165
##  3  1995 F        171
##  4  1995 F         NA
##  5  1995 M         41
##  6  1995 F         45
##  7  1995 M         46
##  8  1995 F         49
##  9  1995 M         46
## 10  1995 M         48
## # ... with 1,170 more rows
```

```
surveys %>%
  mutate(weight_kg = weight/1000,  #creates new column
         weight_kg2 = weight_kg*2)  # creates new column based on new column
```

```
## # A tibble: 34,786 x 15
##    record_id month   day  year plot_id species_id sex   hindfoot_length
##        <dbl> <dbl> <dbl> <dbl>   <dbl> <chr>      <chr>           <dbl>
##  1         1     7    16  1977       2 NL         M                  32
##  2        72     8    19  1977       2 NL         M                  31
##  3       224     9    13  1977       2 NL         <NA>               NA
##  4       266    10    16  1977       2 NL         <NA>               NA
##  5       349    11    12  1977       2 NL         <NA>               NA
##  6       363    11    12  1977       2 NL         <NA>               NA
##  7       435    12    10  1977       2 NL         <NA>               NA
##  8       506     1     8  1978       2 NL         <NA>               NA
##  9       588     2    18  1978       2 NL         M                  NA
## 10       661     3    11  1978       2 NL         <NA>               NA
## # ... with 34,776 more rows, and 7 more variables: weight <dbl>,
## #   genus <chr>, species <chr>, taxa <chr>, plot_type <chr>,
## #   weight_kg <dbl>, weight_kg2 <dbl>
```

```
surveys %>%
  drop_na(weight)  # drops NAs from specified column
```

```
## # A tibble: 32,283 x 13
##    record_id month   day  year plot_id species_id sex   hindfoot_length
##        <dbl> <dbl> <dbl> <dbl>   <dbl> <chr>      <chr>           <dbl>
##  1       588     2    18  1978       2 NL         M                  NA
##  2       845     5     6  1978       2 NL         M                  32
##  3       990     6     9  1978       2 NL         M                  NA
##  4      1164     8     5  1978       2 NL         M                  34
```

```
##  5      1261     9     4  1978       2 NL         M                 32
##  6      1453    11     5  1978       2 NL         M                 NA
##  7      1756     4    29  1979       2 NL         M                 33
##  8      1818     5    30  1979       2 NL         M                 32
##  9      1882     7     4  1979       2 NL         M                 32
## 10      2133    10    25  1979       2 NL         F                 33
## # ... with 32,273 more rows, and 5 more variables: weight <dbl>,
## #   genus <chr>, species <chr>, taxa <chr>, plot_type <chr>
```

Challenge 2

Only species_id. New column called hindfoot_half: contains half foot length values. Also, no NAs in hindfoot_half column. Values <30

```
surveys_jlm<-surveys %>%
  drop_na(hindfoot_length) %>%
  mutate(hindfoot_half=hindfoot_length/2) %>%
  filter(hindfoot_half <30) %>%
  select(species_id, hindfoot_half)
```

```
surveys %>%     #find mean weight by sex and ignore NAs in weight
  group_by(sex) %>%
  summarize(mean_weight = mean(weight, na.rm = TRUE))
```

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

```
surveys %>%     #find mean weight by species and sex and ignore NAs in weight column
  group_by(sex, species_id) %>%
  summarize(mean_weight = mean(weight, na.rm = TRUE))
```

```
## # A tibble: 92 x 3
## # Groups:   sex [?]
##    sex   species_id mean_weight
##    <chr> <chr>            <dbl>
##  1 F     BA                9.16
##  2 F     DM               41.6
##  3 F     DO               48.5
##  4 F     DS              118.
##  5 F     NL              154.
##  6 F     OL               31.1
##  7 F     OT               24.8
##  8 F     OX               21
##  9 F     PB               30.2
## 10 F     PE               22.8
## # ... with 82 more rows
```

```
surveys %>%     #find mean and min weight by species and sex and ignore NAs in weight column
  group_by(sex, species_id) %>%
  summarize(mean_weight = mean(weight, na.rm = TRUE),
            min_weight = min(weight, na.rm = TRUE))
```

```
## # A tibble: 92 x 4
```

```
## # Groups:   sex [?]
##    sex   species_id mean_weight min_weight
##    <chr> <chr>            <dbl>      <dbl>
##  1 F     BA                9.16          6
##  2 F     DM               41.6          10
##  3 F     DO               48.5          12
##  4 F     DS              118.           45
##  5 F     NL              154.           32
##  6 F     OL               31.1          10
##  7 F     OT               24.8           5
##  8 F     OX               21            20
##  9 F     PB               30.2          12
## 10 F     PE               22.8          11
## # ... with 82 more rows
```

```
surveys %>%     #find mean and min weight by species and sex,ignore NAs in weight column and sort by m
  group_by(sex, species_id) %>%
  summarize(mean_weight = mean(weight, na.rm = TRUE),
            min_weight = min(weight, na.rm = TRUE)) %>%
   arrange(min_weight)        # default is ascending, for descending need to add function arrange(desc(
```

```
## # A tibble: 92 x 4
## # Groups:   sex [3]
##    sex   species_id mean_weight min_weight
##    <chr> <chr>            <dbl>      <dbl>
##  1 F     PF                7.97          4
##  2 F     RM               11.1           4
##  3 M     PF                7.89          4
##  4 M     PP               17.2           4
##  5 M     RM               10.1           4
##  6 <NA>  PF                6             4
##  7 F     OT               24.8           5
##  8 F     PP               17.2           5
##  9 F     BA                9.16          6
## 10 M     BA                7.36          6
## # ... with 82 more rows
```

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

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

Challenge 3

1. How many animals were caught in each plot type surveyed?
2. Use group_by and summarize to find mean, min & max of hindfoot_length for each species (using species_id). Also, add no. observations (hint: see ?n)
3. What was the heaviest animal measured in each year? Return year, genus, species_id and weight

```
#1.

surveys %>%
  count(plot_type)
```

```
## # 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
```

```
#2.

surveys %>%
  group_by(species_id) %>%
    summarize(mean_length = mean(hindfoot_length, na.rm = TRUE),
              min_length = min(hindfoot_length, na.rm = TRUE),
              max_length = max(hindfoot_length, na.rm = TRUE),
              n = n())
```

```
## # A tibble: 48 x 5
##    species_id mean_length min_length max_length     n
##    <chr>            <dbl>      <dbl>      <dbl> <int>
##  1 AB                 NaN        Inf       -Inf   303
##  2 AH                  33         31         35   437
##  3 AS                 NaN        Inf       -Inf     2
##  4 BA                  13          6         16    46
##  5 CB                 NaN        Inf       -Inf    50
##  6 CM                 NaN        Inf       -Inf    13
##  7 CQ                 NaN        Inf       -Inf    16
##  8 CS                 NaN        Inf       -Inf     1
##  9 CT                 NaN        Inf       -Inf     1
## 10 CU                 NaN        Inf       -Inf     1
## # ... with 38 more rows
```

```
#3.


max_weights <-surveys %>%
  drop_na(weight) %>%
  group_by(year) %>%
  filter(weight==max(weight)) %>%
  select(year, genus, species, weight) %>%
  arrange(year)

max_weights
```

```
## # A tibble: 27 x 4
## # Groups:   year [26]
##     year genus     species     weight
##    <dbl> <chr>     <chr>        <dbl>
##  1  1977 Dipodomys spectabilis    149
```

```
##  2  1978 Neotoma   albigula       232
##  3  1978 Neotoma   albigula       232
##  4  1979 Neotoma   albigula       274
##  5  1980 Neotoma   albigula       243
##  6  1981 Neotoma   albigula       264
##  7  1982 Neotoma   albigula       252
##  8  1983 Neotoma   albigula       256
##  9  1984 Neotoma   albigula       259
## 10  1985 Neotoma   albigula       225
## # ... with 17 more rows
```

Export our data

```
write_csv(max_weights, here("write_data", "max_weights.csv"))
```