# r workshop

*David Costalago*

*21 de enero de 2019*

### Creating objects in r

```
3 + 5
```

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

```
weight_kg <- 55
```

## Vectors and data types

This section describes some basic data types in r

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

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

### Data Frames

Next we look at the structure of Data Frames

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.4.4
```

```
## -- Attaching packages ----------------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.0.0     v purrr   0.2.4
## v tibble  1.4.1     v dplyr   0.7.6
## v tidyr   0.8.1     v stringr 1.2.0
## v readr   1.1.1     v forcats 0.3.0
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```
## Warning: package 'tidyr' was built under R version 3.4.4
```

```
## Warning: package 'dplyr' was built under R version 3.4.4
```

```
## Warning: package 'forcats' was built under R version 3.4.4
```

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

```
#utils::download.file() #this forces r to use the donwload.file function within the package ´utils´

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 easy
```

```
## Warning: package 'here' was built under R version 3.4.4
```

```
## here() starts at C:/Users/Ordenador/Documents/SPERA_MITACS_JuvenileSalmonProgram/Workshop R for PELa
```

```
surveys <- read_csv(here("read_data", "portal_data_joined.csv"))
```

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

```
surveys
```

```
## # A tibble: 34,786 x 13
##    reco~ month   day  year plot~ spec~ sex   hind~ weig~ genus spec~ taxa
##    <int> <int> <int> <int> <int> <chr> <chr> <int> <int> <chr> <chr> <chr>
##  1     1     7    16  1977     2 NL    M        32    NA Neot~ albi~ Rode~
##  2    72     8    19  1977     2 NL    M        31    NA Neot~ albi~ Rode~
##  3   224     9    13  1977     2 NL    <NA>     NA    NA Neot~ albi~ Rode~
##  4   266    10    16  1977     2 NL    <NA>     NA    NA Neot~ albi~ Rode~
##  5   349    11    12  1977     2 NL    <NA>     NA    NA Neot~ albi~ Rode~
##  6   363    11    12  1977     2 NL    <NA>     NA    NA Neot~ albi~ Rode~
##  7   435    12    10  1977     2 NL    <NA>     NA    NA Neot~ albi~ Rode~
##  8   506     1     8  1978     2 NL    <NA>     NA    NA Neot~ albi~ Rode~
##  9   588     2    18  1978     2 NL    M        NA   218 Neot~ albi~ Rode~
## 10   661     3    11  1978     2 NL    <NA>     NA    NA Neot~ albi~ Rode~
## # ... with 34,776 more rows, and 1 more variable: plot_type <chr>
```

```
dim(surveys)
```

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

```
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 lets use square bracket subsetting.

```
# First define the row coordinate, and then the column.
surveys[1,1]
```

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

```
# Defining only which column we want  will return a data frame
surveys[1]
```

```
## # A tibble: 34,786 x 1
##    record_id
##        <int>
##  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] #gives rows 1 to 3 in column 7
```

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

```
surveys[,-7] #all the rows and columns except column 7
```

```
## # A tibble: 34,786 x 12
##    reco~ month   day  year plot~ spec~ hind~ weig~ genus spec~ taxa  plot~
##    <int> <int> <int> <int> <int> <chr> <int> <int> <chr> <chr> <chr> <chr>
##  1     1     7    16  1977     2 NL       32    NA Neot~ albi~ Rode~ Cont~
##  2    72     8    19  1977     2 NL       31    NA Neot~ albi~ Rode~ Cont~
##  3   224     9    13  1977     2 NL       NA    NA Neot~ albi~ Rode~ Cont~
##  4   266    10    16  1977     2 NL       NA    NA Neot~ albi~ Rode~ Cont~
##  5   349    11    12  1977     2 NL       NA    NA Neot~ albi~ Rode~ Cont~
##  6   363    11    12  1977     2 NL       NA    NA Neot~ albi~ Rode~ Cont~
##  7   435    12    10  1977     2 NL       NA    NA Neot~ albi~ Rode~ Cont~
##  8   506     1     8  1978     2 NL       NA    NA Neot~ albi~ Rode~ Cont~
##  9   588     2    18  1978     2 NL       NA   218 Neot~ albi~ Rode~ Cont~
## 10   661     3    11  1978     2 NL       NA    NA Neot~ albi~ Rode~ Cont~
## # ... with 34,776 more rows
```

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

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

### Data manipulation

Square brackets are good when the coordinates of your data frame are fixed. But what happens if the coordinates change (e.g. you created new columns)

Key functions for data manipulation in dplyr:

- select(): subsetting columns
- filter(): subsets of rows based on condiitons
- mutate(): create new columns, nased on information from other columns
- group_by(): creates groups based on categorical data in a column
- summarize(): creates 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
##      <int> <chr>       <int>
##  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
```

```
#Negative subsetting
select(surveys, -record_id)
```

```
## # A tibble: 34,786 x 12
##    month   day  year plot~ spec~ sex   hind~ weig~ genus spec~ taxa  plot~
##    <int> <int> <int> <int> <chr> <chr> <int> <int> <chr> <chr> <chr> <chr>
##  1     7    16  1977     2 NL    M        32    NA Neot~ albi~ Rode~ Cont~
##  2     8    19  1977     2 NL    M        31    NA Neot~ albi~ Rode~ Cont~
##  3     9    13  1977     2 NL    <NA>     NA    NA Neot~ albi~ Rode~ Cont~
##  4    10    16  1977     2 NL    <NA>     NA    NA Neot~ albi~ Rode~ Cont~
##  5    11    12  1977     2 NL    <NA>     NA    NA Neot~ albi~ Rode~ Cont~
##  6    11    12  1977     2 NL    <NA>     NA    NA Neot~ albi~ Rode~ Cont~
##  7    12    10  1977     2 NL    <NA>     NA    NA Neot~ albi~ Rode~ Cont~
##  8     1     8  1978     2 NL    <NA>     NA    NA Neot~ albi~ Rode~ Cont~
##  9     2    18  1978     2 NL    M        NA   218 Neot~ albi~ Rode~ Cont~
## 10     3    11  1978     2 NL    <NA>     NA    NA Neot~ albi~ Rode~ Cont~
## # ... with 34,776 more rows
```

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

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

```
## # A tibble: 8 x 13
##   recor~ month   day  year plot~ spec~ sex   hind~ weig~ genus spec~ taxa
##    <int> <int> <int> <int> <int> <chr> <chr> <int> <int> <chr> <chr> <chr>
## 1  22314     6     7  1995     2 NL    M        34    NA Neot~ albi~ Rode~
## 2  22728     9    23  1995     2 NL    F        32   165 Neot~ albi~ Rode~
## 3  22899    10    28  1995     2 NL    F        32   171 Neot~ albi~ Rode~
## 4  23032    12     2  1995     2 NL    F        33    NA Neot~ albi~ Rode~
## 5  22847    10    28  1995    12 NL    M        34   138 Neot~ albi~ Rode~
## 6  22998    12     2  1995    12 NL    M        33   152 Neot~ albi~ Rode~
## 7  23124    12    21  1995    12 NL    F        32   160 Neot~ albi~ Rode~
## 8  22476     7    20  1995    24 NL    F        31   149 Neot~ albi~ Rode~
## # ... with 1 more variable: plot_type <chr>
```

## Pipes

Pipes allow you yo chain together dplyr functions.

%>% or ctrl-shift-m

```
# Write multiple arguments in a sentence using pipes
surveys %>%
  filter(weight < 5) %>%
  select(species_id, sex, weight)
```

```
## # A tibble: 17 x 3
##    species_id sex   weight
##    <chr>      <chr>  <int>
##  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

Using pipes, subset the surveys dataframe to include animals collected before 1995 and retain only the columns year, sex and weight.

```
surveys_animal_pre1995 <- surveys %>%
   filter(year < 1995) %>%
  select(year, sex, weight)
```

```
surveys %>%
  mutate(weight_kg = weight / 1000,
         weight_kg2 = weight_kg * 2)
```

```
## # A tibble: 34,786 x 15
##    reco~ month   day  year plot~ spec~ sex   hind~ weig~ genus spec~ taxa
##    <int> <int> <int> <int> <int> <chr> <chr> <int> <int> <chr> <chr> <chr>
##  1     1     7    16  1977     2 NL    M        32    NA Neot~ albi~ Rode~
##  2    72     8    19  1977     2 NL    M        31    NA Neot~ albi~ Rode~
##  3   224     9    13  1977     2 NL    <NA>     NA    NA Neot~ albi~ Rode~
##  4   266    10    16  1977     2 NL    <NA>     NA    NA Neot~ albi~ Rode~
##  5   349    11    12  1977     2 NL    <NA>     NA    NA Neot~ albi~ Rode~
##  6   363    11    12  1977     2 NL    <NA>     NA    NA Neot~ albi~ Rode~
##  7   435    12    10  1977     2 NL    <NA>     NA    NA Neot~ albi~ Rode~
##  8   506     1     8  1978     2 NL    <NA>     NA    NA Neot~ albi~ Rode~
##  9   588     2    18  1978     2 NL    M        NA   218 Neot~ albi~ Rode~
## 10   661     3    11  1978     2 NL    <NA>     NA    NA Neot~ albi~ Rode~
## # ... with 34,776 more rows, and 3 more variables: plot_type <chr>,
## #   weight_kg <dbl>, weight_kg2 <dbl>
```

```
surveys <- surveys %>%
  drop_na(weight) %>%
  mutate(mean_weight = mean(weight))
```

```
mean(surveys$weight)
```

```
## [1] 42.67243
```

Challenge #2

Contains only the species_id column, has a new column called hindfoot_half that are half the hindfoot_length values. Also, in the new hinfoot_half column there are no NAs and values are less than 30

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

```
surveys %>%
  group_by(sex, species_id) %>%
  summarize(mean_weight = mean(weight, na.rm = TRUE),
            min_weight = min(weight, na.rm = TRUE)) %>%
  arrange(min_weight) # arrange is done by ascending order by default
```

```
## # A tibble: 64 x 4
## # Groups: sex [3]
##    sex   species_id mean_weight min_weight
##    <chr> <chr>            <dbl>      <dbl>
##  1 F     PF                7.97       4.00
##  2 F     RM               11.1        4.00
##  3 M     PF                7.89       4.00
##  4 M     PP               17.2        4.00
##  5 M     RM               10.1        4.00
##  6 <NA>  PF                6.00       4.00
##  7 F     OT               24.8        5.00
##  8 F     PP               17.2        5.00
##  9 F     BA                9.16       6.00
## 10 M     BA                7.36       6.00
## # ... with 54 more rows
```

```
surveys %>%
  count(sex) #handy to get samples sizes for different groups
```

```
## # A tibble: 3 x 2
##   sex       n
##   <chr> <int>
## 1 F     15303
## 2 M     16879
## 3 <NA>    101
```

```
# the above code is synonomous
surveys %>%
  group_by(sex) %>%
  summarize(count = n ())
```

```
## # A tibble: 3 x 2
##   sex   count
##   <chr> <int>
## 1 F     15303
## 2 M     16879
```

```
## 3 <NA>    101
```

Challenge #3

1. How many animals were caught in each plot_type surveyed.
2. Use group_by and summarize to find the mean, min and max of hindfoot length (using species_id) for each species. Also, add the number of obervacions (hint: see ?n).
3. What was the heaviest animal measured in each year? Return the columns year, genus, species_id, and weight

```
#1
surveys %>%
  count(plot_type)
```

```
## # A tibble: 5 x 2
##   plot_type                     n
##   <chr>                     <int>
## 1 Control                   14652
## 2 Long-term Krat Exclosure   4692
## 3 Rodent Exclosure           3818
## 4 Short-term Krat Exclosure  5407
## 5 Spectab exclosure          3714
```

```
#2
surveys %>%
  group_by(species_id) %>%
  summarize(count = n (),
            mean(hindfoot_length, na.rm = TRUE),
            max(hindfoot_length, na.rm = TRUE),
            min(hindfoot_length, na.rm = TRUE))
```

```
## # A tibble: 25 x 5
##    species_id count `mean(hindfoot_length, na.rm = TRUE)` `max(hi~ `min(h~
##    <chr>      <int>                                 <dbl>    <dbl>   <dbl>
##  1 BA            45                                  13.0     16.0    6.00
##  2 DM         10262                                  36.0     50.0   16.0
##  3 DO          2904                                  35.6     64.0   26.0
##  4 DS          2344                                  50.0     58.0   39.0
##  5 NL          1152                                  32.3     42.0   21.0
##  6 OL           970                                  20.5     39.0   12.0
##  7 OT          2160                                  20.3     50.0   13.0
##  8 OX             6                                  20.4     21.0   19.0
##  9 PB          2810                                  26.1     47.0    2.00
## 10 PE          1260                                  20.2     30.0   11.0
## # ... with 15 more rows
```

```
#3
surveys %>%
  group_by(year, genus, species_id) %>%
  summarize(max_weight = max(weight, na.rm = TRUE)) %>%
  arrange(desc(max_weight))
```

```
## # A tibble: 336 x 4
## # Groups: year, genus [201]
##     year genus   species_id max_weight
##    <int> <chr>   <chr>           <dbl>
```

```
##  1  2001 Neotoma NL                280
##  2  1987 Neotoma NL                278
##  3  1989 Neotoma NL                275
##  4  1979 Neotoma NL                274
##  5  2000 Neotoma NL                265
##  6  1981 Neotoma NL                264
##  7  1984 Neotoma NL                259
##  8  1983 Neotoma NL                256
##  9  1982 Neotoma NL                252
## 10  1988 Neotoma NL                248
## # ... with 326 more rows
```

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

### Export our data

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