SIOB 296 Introduction to Programming with R

Eric Archer (eric.archer@noaa.gov)

Week 10: March 9, 2020

File/Folder Management

list files in a folder

The `dir()` function lists all of the files in a folder. Keep in mind that it returns a character vector that can be saved to an object to be used later. The `pattern` argument, lists only files that match the specified pattern. Setting `full.names = TRUE` will return the full path of the files. Setting `recursive = TRUE` will provide a list to all subdirectories.

```
# here's a full listing of all .rdata files in the parent folder
files <- dir("..", pattern = ".rdata", full.names = TRUE, recursive = TRUE)
head(files)
```

```
[1] "../Prep/Data 02.rdata"         "../Prep/eye.color.rdata"
[3] "../Prep/merge data.rdata"      "../Prep/test ws.rdata"
[5] "../Prep/xy.rdata"              "../Week 01/both objects.rdata"
```

We can test if a file or folder is present with `file.exists()`:

```
dir()
```

```
 [1] "coords.csv"                "ctd 2012.R"
 [3] "ctd data"                  "ctd positiions.csv"
 [5] "ctd.csv"                   "Data 02.R"
 [7] "Data 02.rdata"             "Data Structures.jpg"
 [9] "extract 33 ctd stations.R" "extract ctd data.R"
[11] "eye.color.rdata"           "free text.txt"
[13] "Indexing.jpg"              "lm.R"
[15] "merge data.rdata"          "multiYearCTD.csv"
[17] "regression example.R"      "tblCodeSpecies.csv"
[19] "temperature.r"             "test ws.rdata"
[21] "test.csv"                  "Week 01 Homework.Rmd"
[23] "Week 01 Notes.Rmd"         "Week 02 Homework.Rmd"
[25] "Week 02 Notes.Rmd"         "Week 04 Homework.Rmd"
[27] "Week 04 Notes.Rmd"         "Week 05 Homework.Rmd"
[29] "Week 05 Notes.Rmd"         "Week 06 Homework.Rmd"
[31] "Week 06 Notes.Rmd"         "Week 08 Homework.Rmd"
[33] "Week 08 Notes.Rmd"         "Week 09 Notes.Rmd"
[35] "Week 10 Notes.Rmd"         "Week-02-Homework.pdf"
[37] "Week-02-Notes.pdf"         "Week-04-Homework.pdf"
[39] "Week-04-Notes.pdf"         "Week-05-Homework.pdf"
[41] "Week-05-Notes.pdf"         "Week-10-Notes.pdf"
[43] "Week-10-Notes.Rmd"         "x.r"
[45] "xy.rdata"
```

```
file.exists("missing.rdata")
```

```
[1] FALSE
```

```
x <- 1
save(x, file = "x test.rdata")
file.exists("x test.rdata")
```

```
[1] TRUE
```

To delete a file, use `file.remove()`:

```
file.remove("x test.rdata")
```

```
[1] TRUE
```

```
file.exists("x test.rdata")
```

```
[1] FALSE
```

You can't delete a directory that is not empty with `file.remove()`. For this, you need to use `unlink()`. You should include the `recursive = TRUE` argument to delete all files and subdirectories contained in the directory being deleted:

```
unlink("new dir", recursive = TRUE)
dir.exists("new dir")
```

```
[1] FALSE
```

We can create a new directory with `dir.create()`:

```
dir.create("new dir")
dir()
```

```
 [1] "coords.csv"                "ctd 2012.R"
 [3] "ctd data"                  "ctd positiions.csv"
 [5] "ctd.csv"                   "Data 02.R"
 [7] "Data 02.rdata"             "Data Structures.jpg"
 [9] "extract 33 ctd stations.R" "extract ctd data.R"
[11] "eye.color.rdata"           "free text.txt"
[13] "Indexing.jpg"              "lm.R"
[15] "merge data.rdata"          "multiYearCTD.csv"
[17] "new dir"                   "regression example.R"
[19] "tblCodeSpecies.csv"        "temperature.r"
[21] "test ws.rdata"             "test.csv"
[23] "Week 01 Homework.Rmd"      "Week 01 Notes.Rmd"
[25] "Week 02 Homework.Rmd"      "Week 02 Notes.Rmd"
[27] "Week 04 Homework.Rmd"      "Week 04 Notes.Rmd"
[29] "Week 05 Homework.Rmd"      "Week 05 Notes.Rmd"
[31] "Week 06 Homework.Rmd"      "Week 06 Notes.Rmd"
[33] "Week 08 Homework.Rmd"      "Week 08 Notes.Rmd"
[35] "Week 09 Notes.Rmd"         "Week 10 Notes.Rmd"
[37] "Week-02-Homework.pdf"      "Week-02-Notes.pdf"
[39] "Week-04-Homework.pdf"      "Week-04-Notes.pdf"
[41] "Week-05-Homework.pdf"      "Week-05-Notes.pdf"
[43] "Week-10-Notes.pdf"         "Week-10-Notes.Rmd"
[45] "x.r"                       "xy.rdata"
```

In order to create paths to files that are correct regardless of the OS you're using, use the `file.path()` function:

```
x <- 1
x.fname <- file.path("new dir", "x ws.rdata")
x.fname
```

```
[1] "new dir/x ws.rdata"
```

```
save(x, file = x.fname)
dir("new dir", full.names = TRUE)
```

```
[1] "new dir/x ws.rdata"
```

To remove all path specifications of a filename, use `basename()`:

```
rdata.files <- dir("..", pattern = ".rdata", recursive = TRUE)
head(rdata.files)
```

```
[1] "Prep/Data 02.rdata"      "Prep/eye.color.rdata"
[3] "Prep/merge data.rdata"   "Prep/new dir/x ws.rdata"
[5] "Prep/test ws.rdata"      "Prep/xy.rdata"
```

```
head(basename(rdata.files))
```

```
[1] "Data 02.rdata"    "eye.color.rdata"  "merge data.rdata" "x ws.rdata"
[5] "test ws.rdata"    "xy.rdata"
```

The reverse, `dirname()` returns just the path portion:

```
dirname(rdata.files)
```

```
 [1] "Prep"         "Prep"         "Prep"         "Prep/new dir" "Prep"
 [6] "Prep"         "Week 01"      "Week 01"      "Week 01"      "Week 01"
[11] "Week 02"      "Week 09"
```

Finding files of a particular extension requires the use of regular expressions. The regular expression that we need is "\.ext$", which matches all strings with ".ext" at the end. So, to find all ".csv" files, we use:

```
csv.fnames <- dir("..", pattern = "\\.csv$", full.names = TRUE, recursive = TRUE)
head(csv.fnames, 5)
```

```
[1] "../Prep/coords.csv"
[2] "../Prep/ctd data/ctd positions.csv"
[3] "../Prep/ctd data/Station.1 2010-03-01.csv"
[4] "../Prep/ctd data/Station.1 2010-05-04.csv"
[5] "../Prep/ctd data/Station.1 2010-08-11.csv"
```

To remove the extension, we use the same string with the `gsub()` function:

```
csv.fnames <- gsub("\\.csv$", "", csv.fnames)
head(basename(csv.fnames), 5)
```

```
[1] "coords"               "ctd positions"        "Station.1 2010-03-01"
[4] "Station.1 2010-05-04" "Station.1 2010-08-11"
```

Piping

```
# method one of doing three steps (sequential)
x <- runif(100)
x.q <- quantile(x, c(0.025, 0.975))
x.q.diff.1 <- diff(x.q)
```

```
# method two (nested)
x.q.diff.2 <- diff(quantile(runif(100), c(0.025, 0.975)))
```

Piping (from package `magrittr`) uses the `%>%` operator

```
library(magrittr)
runif(10)
```

```
 [1] 0.95894943 0.91640423 0.66769056 0.67474486 0.05198777 0.34717763
 [7] 0.24417347 0.72948728 0.77843700 0.62441081
```

```
10 %>% runif()
```

```
 [1] 0.6374459 0.8595522 0.7304642 0.4423203 0.9243625 0.9847265 0.4858135
 [8] 0.4992046 0.9497783 0.4178909
```

```
# no parentheses needed if left side is all that is going into function
10 %>% runif
```

```
 [1] 0.2560267 0.7679528 0.7347302 0.8319930 0.2757561 0.3859313 0.1870058
 [8] 0.2891994 0.1357705 0.8188193
```

```
# using arguments
10 %>% runif(100, 200)
```

```
 [1] 193.9733 199.6521 114.8349 106.1324 109.8581 193.5835 146.2957 171.4165
 [9] 194.0422 163.9099
```

```
# pipe to second argument (must name arguments)
100 %>% runif(n = 5, max = 200)
```

```
[1] 154.3933 176.2421 189.7173 148.5283 131.2565
```

```
# vs...
100 %>% runif(5, 200)
```

```
  [1] 125.597596 130.108535 100.394877  97.603682  56.181408  47.686342
  [7]  33.237708  28.040069 135.151097 192.743860   7.936728  98.847737
 [13] 163.092538 142.586764 113.607161 115.579724  41.201404  64.064309
 [19]  31.717763 119.054291 148.244570 111.076568 160.640626  89.180462
 [25]   7.030829  53.906374 129.130453 185.571804  38.194180  31.181775
 [31] 180.787409 140.277360 134.929622  65.053032 176.502800 159.113458
 [37] 123.853052 103.288960 132.631821 145.861864 114.045239  38.893520
 [43] 151.515388  16.067200 194.542912 181.453230  63.384997  47.091431
 [49] 161.742038 143.371995  20.348857 187.414054 185.577822 105.762111
 [55]  10.699301 185.941994 151.837120 129.431924 157.169781 160.410912
 [61] 125.754977  27.207747  17.275730  29.761283 104.196958 162.710329
 [67]  57.814447  28.560075 105.459745  26.724534 151.936294 148.206873
 [73]  31.077463 120.492986   5.031756 183.920355  78.640362 187.006791
 [79] 129.916198  86.135275 120.798994 133.477117 152.055932 103.295495
 [85] 140.035459 128.479311 169.176670  32.158622  43.346683  35.676803
 [91]  43.396766 186.055949 189.245505 194.842101  94.932713 128.803947
 [97] 146.479028  97.240071  13.571170 153.569821
```

pipe version of first example

```
q.diff.pipe <- 100 %>%
  runif %>%
  quantile(c(0.025, 0.975)) %>%
  diff
```

dplyr

The Data Wrangling Cheatsheet: https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf

filter and select

```
library(tidyverse)
```

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

```
v ggplot2 3.2.1     v purrr   0.3.3
v tibble  2.1.3     v dplyr   0.8.4
v tidyr   1.0.2     v stringr 1.4.0
v readr   1.3.1     v forcats 0.4.0
```

```
-- Conflicts --------------------- tidyverse_conflicts() --
x tidyr::extract()   masks magrittr::extract()
x dplyr::filter()    masks stats::filter()
x dplyr::lag()       masks stats::lag()
x purrr::set_names() masks magrittr::set_names()
```

```
# base R indexing to select males
#starwars[starwars$gender == "male", ]
#subset(starwars, gender == "male")

# dplyr way - filter
filter(starwars, gender == "male")
```

```
# A tibble: 62 x 13
   name  height  mass hair_color skin_color eye_color birth_year gender
   <chr>  <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr>
 1 Luke~    172    77 blond      fair       blue            19   male
 2 Dart~    202   136 none       white      yellow          41.9 male
 3 Owen~    178   120 brown, gr~ light      blue            52   male
 4 Bigg~    183    84 black      light      brown           24   male
 5 Obi-~    182    77 auburn, w~ fair       blue-gray       57   male
 6 Anak~    188    84 blond      fair       blue            41.9 male
 7 Wilh~    180    NA auburn, g~ fair       blue            64   male
 8 Chew~    228   112 brown      unknown    blue           200   male
 9 Han ~    180    80 brown      fair       brown           29   male
10 Gree~    173    74 <NA>       green      black           44   male
# ... with 52 more rows, and 5 more variables: homeworld <chr>, species <chr>,
#   films <list>, vehicles <list>, starships <list>
```

```
# pipeline version
starwars %>%
  filter(gender == "male" & height > 190)
```

```
# A tibble: 20 x 13
   name  height  mass hair_color skin_color eye_color birth_year gender
   <chr>  <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr>
 1 Dart~    202   136 none       white      yellow          41.9 male
 2 Chew~    228   112 brown      unknown    blue           200   male
 3 Qui-~    193    89 brown      fair       blue            92   male
 4 Nute~    191    90 none       mottled g~ red             NA   male
 5 Jar ~    196    66 none       orange     orange          52   male
 6 Roos~    224    82 none       grey       orange          NA   male
```

```
 7 Rugo~    206    NA none       green      orange          NA   male
 8 Ki-A~    198    82 white      pale       yellow          92   male
 9 Kit ~    196    87 none       green      black           NA   male
10 Yara~    264    NA none       white      yellow          NA   male
11 Mas ~    196    NA none       blue       blue            NA   male
12 Dooku    193    80 white      fair       brown          102   male
13 Bail~    191    NA black      tan        brown           67   male
14 Dext~    198   102 none       brown      yellow          NA   male
15 Lama~    229    88 none       grey       black           NA   male
16 Wat ~    193    48 none       green, gr~ unknown         NA   male
17 San ~    191    NA none       grey       gold            NA   male
18 Grie~    216   159 none       brown, wh~ green, y~       NA   male
19 Tarf~    234   136 brown      brown      blue            NA   male
20 Tion~    206    80 none       grey       black           NA   male
# ... with 5 more variables: homeworld <chr>, species <chr>, films <list>,
#   vehicles <list>, starships <list>
```

```
# "select" columns to return
select(starwars, name, height, mass, gender)
```

```
# A tibble: 87 x 4
   name               height  mass gender
   <chr>               <int> <dbl> <chr>
 1 Luke Skywalker        172    77 male
 2 C-3PO                 167    75 <NA>
 3 R2-D2                  96    32 <NA>
 4 Darth Vader           202   136 male
 5 Leia Organa           150    49 female
 6 Owen Lars             178   120 male
 7 Beru Whitesun lars    165    75 female
 8 R5-D4                  97    32 <NA>
 9 Biggs Darklighter     183    84 male
10 Obi-Wan Kenobi        182    77 male
# ... with 77 more rows
```

```
select(starwars, height, gender, name, mass)
```

```
# A tibble: 87 x 4
   height gender name               mass
    <int> <chr>  <chr>             <dbl>
 1    172 male   Luke Skywalker       77
 2    167 <NA>   C-3PO                75
 3     96 <NA>   R2-D2                32
 4    202 male   Darth Vader         136
 5    150 female Leia Organa          49
 6    178 male   Owen Lars           120
 7    165 female Beru Whitesun lars   75
 8     97 <NA>   R5-D4                32
 9    183 male   Biggs Darklighter    84
10    182 male   Obi-Wan Kenobi       77
# ... with 77 more rows
```

```
# extend pipeline above
starwars %>%
  filter(gender == "male" & height > 190) %>%
  select(name, height, mass)
```

```
# A tibble: 20 x 3
   name                height  mass
   <chr>                <int> <dbl>
 1 Darth Vader            202   136
 2 Chewbacca              228   112
 3 Qui-Gon Jinn           193    89
 4 Nute Gunray            191    90
 5 Jar Jar Binks          196    66
 6 Roos Tarpals           224    82
 7 Rugor Nass             206    NA
 8 Ki-Adi-Mundi           198    82
 9 Kit Fisto              196    87
10 Yarael Poof            264    NA
11 Mas Amedda             196    NA
12 Dooku                  193    80
13 Bail Prestor Organa    191    NA
14 Dexter Jettster        198   102
15 Lama Su                229    88
16 Wat Tambor             193    48
17 San Hill               191    NA
18 Grievous               216   159
19 Tarfful                234   136
20 Tion Medon             206    80
```

```r
# helper functions for select

# select range of columns
starwars %>%
  filter(gender == "male" & height > 190) %>%
  select(eye_color:homeworld)
```

```
# A tibble: 20 x 4
   eye_color     birth_year gender homeworld
   <chr>              <dbl> <chr>  <chr>
 1 yellow              41.9 male   Tatooine
 2 blue               200   male   Kashyyyk
 3 blue                92   male   <NA>
 4 red                 NA   male   Cato Neimoidia
 5 orange              52   male   Naboo
 6 orange              NA   male   Naboo
 7 orange              NA   male   Naboo
 8 yellow              92   male   Cerea
 9 black               NA   male   Glee Anselm
10 yellow              NA   male   Quermia
11 blue                NA   male   Champala
12 brown              102   male   Serenno
13 brown               67   male   Alderaan
14 yellow              NA   male   Ojom
15 black               NA   male   Kamino
16 unknown             NA   male   Skako
17 gold                NA   male   Muunilinst
18 green, yellow       NA   male   Kalee
19 blue                NA   male   Kashyyyk
20 black               NA   male   Utapau
```

```r
# select columns that start with string
starwars %>%
  filter(gender == "male" & height > 190) %>%
  select(starts_with("h"))
```

```
# A tibble: 20 x 3
   height hair_color homeworld
    <int> <chr>      <chr>
 1    202 none       Tatooine
 2    228 brown      Kashyyyk
 3    193 brown      <NA>
 4    191 none       Cato Neimoidia
 5    196 none       Naboo
 6    224 none       Naboo
 7    206 none       Naboo
 8    198 white      Cerea
 9    196 none       Glee Anselm
10    264 none       Quermia
11    196 none       Champala
12    193 white      Serenno
13    191 black      Alderaan
14    198 none       Ojom
15    229 none       Kamino
16    193 none       Skako
17    191 none       Muunilinst
18    216 none       Kalee
19    234 brown      Kashyyyk
20    206 none       Utapau
```

```r
# select columns that contain a string
starwars %>%
  filter(gender == "male" & height > 190) %>%
  select(contains("color"))
```

```
# A tibble: 20 x 3
   hair_color skin_color    eye_color
   <chr>      <chr>         <chr>
 1 none       white         yellow
 2 brown      unknown       blue
 3 brown      fair          blue
 4 none       mottled green red
 5 none       orange        orange
 6 none       grey          orange
 7 none       green         orange
 8 white      pale          yellow
 9 none       green         black
10 none       white         yellow
11 none       blue          blue
12 white      fair          brown
13 black      tan           brown
14 none       brown         yellow
15 none       grey          black
16 none       green, grey   unknown
17 none       grey          gold
18 none       brown, white  green, yellow
```

```
19 brown      brown         blue
20 none       grey          black
```

```
# select columns excluding certain ones
starwars %>%
  filter(gender == "male" & height > 190) %>%
  select(-name, -gender, -height)
```

```
# A tibble: 20 x 10
    mass hair_color skin_color eye_color birth_year homeworld species films
   <dbl> <chr>      <chr>      <chr>          <dbl> <chr>     <chr>   <lis>
 1   136 none       white      yellow          41.9 Tatooine  Human   <chr~
 2   112 brown      unknown    blue           200   Kashyyyk  Wookiee <chr~
 3    89 brown      fair       blue            92   <NA>      Human   <chr~
 4    90 none       mottled g~ red             NA   Cato Nei~ Neimod~ <chr~
 5    66 none       orange     orange          52   Naboo     Gungan  <chr~
 6    82 none       grey       orange          NA   Naboo     Gungan  <chr~
 7    NA none       green      orange          NA   Naboo     Gungan  <chr~
 8    82 white      pale       yellow          92   Cerea     Cerean  <chr~
 9    87 none       green      black           NA   Glee Ans~ Nautol~ <chr~
10    NA none       white      yellow          NA   Quermia   Quermi~ <chr~
11    NA none       blue       blue            NA   Champala  Chagri~ <chr~
12    80 white      fair       brown          102   Serenno   Human   <chr~
13    NA black      tan        brown           67   Alderaan  Human   <chr~
14   102 none       brown      yellow          NA   Ojom      Besali~ <chr~
15    88 none       grey       black           NA   Kamino    Kamino~ <chr~
16    48 none       green, gr~ unknown         NA   Skako     Skakoan <chr~
17    NA none       grey       gold            NA   Muunilin~ Muun    <chr~
18   159 none       brown, wh~ green, y~       NA   Kalee     Kaleesh <chr~
19   136 brown      brown      blue            NA   Kashyyyk  Wookiee <chr~
20    80 none       grey       black           NA   Utapau    Pau'an  <chr~
# ... with 2 more variables: vehicles <list>, starships <list>
```

arrange to sort data

```
# base R sorting a data.frame
starwars[order(starwars$species, starwars$height), ]
```

```
# A tibble: 87 x 13
   name  height  mass hair_color skin_color eye_color birth_year gender
   <chr>  <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr>
 1 Ratt~     79    15 none       grey, blue unknown           NA male
 2 Dext~    198   102 none       brown      yellow            NA male
 3 Ki-A~    198    82 white      pale       yellow            92 male
 4 Mas ~    196    NA none       blue       blue              NA male
 5 Zam ~    168    55 blonde     fair, gre~ yellow            NA female
 6 R2-D2     96    32 <NA>       white, bl~ red               33 <NA>
 7 R5-D4     97    32 <NA>       white, red red               NA <NA>
 8 C-3PO    167    75 <NA>       gold       yellow           112 <NA>
 9 IG-88    200   140 none       metal      red               15 none
10 BB8       NA    NA none       none       black             NA none
# ... with 77 more rows, and 5 more variables: homeworld <chr>, species <chr>,
#   films <list>, vehicles <list>, starships <list>
```

```
# arrange
starwars %>%
  arrange(species, desc(height)) %>%
```

```
  select(name, height, species)
```

```
# A tibble: 87 x 3
   name            height species
   <chr>            <int> <chr>
 1 Ratts Tyerell       79 Aleena
 2 Dexter Jettster    198 Besalisk
 3 Ki-Adi-Mundi       198 Cerean
 4 Mas Amedda         196 Chagrian
 5 Zam Wesell         168 Clawdite
 6 IG-88              200 Droid
 7 C-3PO              167 Droid
 8 R5-D4               97 Droid
 9 R2-D2               96 Droid
10 BB8                 NA Droid
# ... with 77 more rows
```

new columns

```
sw <- starwars %>%
  mutate(
    height.m = height / 100,
    bmi = mass / height.m ^ 2
  )

# takes place of
# sw <- starwars
# sw$height.m <- sw$height / 100
# sw$bmi <- sw$mass / sw$height.m ^ 2
```

change name of column

```
sw <- starwars %>%
  rename(handle = "name")
colnames(starwars)
```

```
 [1] "name"       "height"     "mass"       "hair_color" "skin_color"
 [6] "eye_color"  "birth_year" "gender"     "homeworld"  "species"
[11] "films"      "vehicles"   "starships"
```

```
colnames(sw)
```

```
 [1] "handle"     "height"     "mass"       "hair_color" "skin_color"
 [6] "eye_color"  "birth_year" "gender"     "homeworld"  "species"
[11] "films"      "vehicles"   "starships"
```

create new column and drop all others

```
sw <- starwars %>%
  transmute(
    name = name,
    height.m = height / 100,
    bmi = mass / height.m ^ 2
  )
sw
```

```
# A tibble: 87 x 3
   name               height.m   bmi
```

```
   <chr>                <dbl> <dbl>
 1 Luke Skywalker        1.72  26.0
 2 C-3PO                 1.67  26.9
 3 R2-D2                 0.96  34.7
 4 Darth Vader           2.02  33.3
 5 Leia Organa           1.5   21.8
 6 Owen Lars             1.78  37.9
 7 Beru Whitesun lars    1.65  27.5
 8 R5-D4                 0.97  34.0
 9 Biggs Darklighter     1.83  25.1
10 Obi-Wan Kenobi        1.82  23.2
# ... with 77 more rows
```

```
# same as
sw <- starwars %>%
  mutate(
    height.m = height / 100,
    bmi = mass / height.m ^ 2
  ) %>%
  select(height.m, bmi)
sw
```

```
# A tibble: 87 x 2
   height.m   bmi
      <dbl> <dbl>
 1     1.72  26.0
 2     1.67  26.9
 3     0.96  34.7
 4     2.02  33.3
 5     1.5   21.8
 6     1.78  37.9
 7     1.65  27.5
 8     0.97  34.0
 9     1.83  25.1
10     1.82  23.2
# ... with 77 more rows
```

complete data set (no missing data)

```
# in base R
#sw.complete <- starwars[complete.cases(starwars), ]

sw.complete <- starwars %>%
  select(-(films:starships), -mass) %>%
  filter(complete.cases(.))

nrow(starwars)
```

```
[1] 87
```

```
nrow(sw.complete)
```

```
[1] 35
```

```
sw.complete
```

```
# A tibble: 35 x 9
   name  height hair_color skin_color eye_color birth_year gender homeworld
```

```
   <chr>  <int> <chr>      <chr>      <chr>          <dbl> <chr>  <chr>
 1 Luke~    172 blond      fair       blue            19   male   Tatooine
 2 Dart~    202 none       white      yellow          41.9 male   Tatooine
 3 Leia~    150 brown      light      brown           19   female Alderaan
 4 Owen~    178 brown, gr~ light      blue            52   male   Tatooine
 5 Beru~    165 brown      light      blue            47   female Tatooine
 6 Bigg~    183 black      light      brown           24   male   Tatooine
 7 Obi-~    182 auburn, w~ fair       blue-gray       57   male   Stewjon
 8 Anak~    188 blond      fair       blue            41.9 male   Tatooine
 9 Wilh~    180 auburn, g~ fair       blue            64   male   Eriadu
10 Chew~    228 brown      unknown    blue           200   male   Kashyyyk
# ... with 25 more rows, and 1 more variable: species <chr>
```

removing duplicates

```
# what are the observed combinations of gender and species
starwars %>%
  select(gender, species) %>%
  distinct() %>%
  arrange(species, gender)
```

```
# A tibble: 43 x 2
   gender species
   <chr>  <chr>
 1 male   Aleena
 2 male   Besalisk
 3 male   Cerean
 4 male   Chagrian
 5 female Clawdite
 6 none   Droid
 7 <NA>   Droid
 8 male   Dug
 9 male   Ewok
10 male   Geonosian
# ... with 33 more rows
```

select random rows

```
# without replacement
starwars %>%
  sample_n(10)
```

```
# A tibble: 10 x 13
   name  height  mass hair_color skin_color eye_color birth_year gender
   <chr>  <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr>
 1 Watto    137    NA black      blue, grey yellow            NA male
 2 Fini~    170    NA blond      fair       blue              91 male
 3 R2-D2     96    32 <NA>       white, bl~ red               33 <NA>
 4 Nute~    191    90 none       mottled g~ red               NA male
 5 Taun~    213    NA none       grey       black             NA female
 6 Dart~    175    80 none       red        yellow            54 male
 7 Jabb~    175  1358 <NA>       green-tan~ orange           600 herma~
 8 Saes~    188    NA none       pale       orange            NA male
 9 Dooku    193    80 white      fair       brown            102 male
10 Finn      NA    NA black      dark       dark              NA male
# ... with 5 more variables: homeworld <chr>, species <chr>, films <list>,
#   vehicles <list>, starships <list>
```

```r
# with replacement
starwars %>%
  sample_n(10, weight = sample(1:10, nrow(.), replace = T))
```

```
# A tibble: 10 x 13
   name  height  mass hair_color skin_color eye_color birth_year gender
   <chr>  <int> <dbl> <chr>      <chr>      <chr>           <dbl> <chr>
 1 Saes~    188    NA none       pale       orange             NA male
 2 Jabb~    175  1358 <NA>       green-tan~ orange            600 herma~
 3 Sebu~    112    40 none       grey, red  orange             NA male
 4 Shaa~    178    57 none       red, blue~ black              NA female
 5 San ~    191    NA none       grey       gold               NA male
 6 Kit ~    196    87 none       green      black              NA male
 7 Dud ~     94    45 none       blue, grey yellow             NA male
 8 Wedg~    170    77 brown      fair       hazel              21 male
 9 Beru~    165    75 brown      light      blue               47 female
10 Ayla~    178    55 none       blue       hazel              48 female
# ... with 5 more variables: homeworld <chr>, species <chr>, films <list>,
#   vehicles <list>, starships <list>
```

group_by

```r
sw <- starwars %>%
  group_by(species) %>%
  summarize(
    mean.height = mean(height, na.rm = T),
    mean.mass = mean(mass, na.rm = T),
    bmi.mean = mean.mass / (mean.height / 100) ^ 2
  )
sw
```

```
# A tibble: 38 x 4
   species   mean.height mean.mass bmi.mean
   <chr>           <dbl>     <dbl>    <dbl>
 1 Aleena             79        15     24.0
 2 Besalisk          198       102     26.0
 3 Cerean            198        82     20.9
 4 Chagrian          196       NaN    NaN
 5 Clawdite          168        55     19.5
 6 Droid             140        69.8   35.6
 7 Dug               112        40     31.9
 8 Ewok               88        20     25.8
 9 Geonosian         183        80     23.9
10 Gungan            209.       74     17.0
# ... with 28 more rows
```

```r
sw <- starwars %>%
  group_by(species, gender) %>%
  summarize(
    mean.height = mean(height, na.rm = T),
    mean.mass = mean(mass, na.rm = T),
    bmi.mean = mean.mass / (mean.height / 100) ^ 2
  )
sw
```

```
# A tibble: 43 x 5
```

```
# Groups:   species [38]
   species   gender mean.height mean.mass bmi.mean
   <chr>     <chr>        <dbl>     <dbl>    <dbl>
 1 Aleena    male            79      15       24.0
 2 Besalisk  male           198     102       26.0
 3 Cerean    male           198      82       20.9
 4 Chagrian  male           196     NaN      NaN
 5 Clawdite  female         168      55       19.5
 6 Droid     none           200     140       35
 7 Droid     <NA>           120      46.3     32.2
 8 Dug       male           112      40       31.9
 9 Ewok      male            88      20       25.8
10 Geonosian male           183      80       23.9
# ... with 33 more rows
```

```
# same summaries, but with mutate on grouped tibble
sw <- starwars %>%
  group_by(species, gender) %>%
  mutate(
    mean.height = mean(height, na.rm = T),
    mean.mass = mean(mass, na.rm = T),
    bmi.mean = mean.mass / (mean.height / 100) ^ 2
  )
sw
```

```
# A tibble: 87 x 16
# Groups:   species, gender [43]
   name  height  mass hair_color skin_color eye_color birth_year gender
   <chr>  <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr>
 1 Luke~    172    77 blond      fair       blue            19   male
 2 C-3PO    167    75 <NA>       gold       yellow         112   <NA>
 3 R2-D2     96    32 <NA>       white, bl~ red             33   <NA>
 4 Dart~    202   136 none       white      yellow          41.9 male
 5 Leia~    150    49 brown      light      brown           19   female
 6 Owen~    178   120 brown, gr~ light      blue            52   male
 7 Beru~    165    75 brown      light      blue            47   female
 8 R5-D4     97    32 <NA>       white, red red             NA   <NA>
 9 Bigg~    183    84 black      light      brown           24   male
10 Obi-~    182    77 auburn, w~ fair       blue-gray       57   male
# ... with 77 more rows, and 8 more variables: homeworld <chr>, species <chr>,
#   films <list>, vehicles <list>, starships <list>, mean.height <dbl>,
#   mean.mass <dbl>, bmi.mean <dbl>
```

```
# same summaries, but with mutate on grouped tibble
sw <- starwars %>%
  group_by(species, gender) %>%
  mutate(
    mean.height = mean(height, na.rm = T),
    mean.mass = mean(mass, na.rm = T),
    bmi.mean = mean.mass / (mean.height / 100) ^ 2,
    bmi = mass / (height / 100) ^ 2
  )
sw
```

```
# A tibble: 87 x 17
# Groups:   species, gender [43]
```

```
   name  height  mass hair_color skin_color eye_color birth_year gender
   <chr>  <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr>
 1 Luke~    172    77 blond      fair       blue            19   male
 2 C-3PO    167    75 <NA>       gold       yellow         112   <NA>
 3 R2-D2     96    32 <NA>       white, bl~ red             33   <NA>
 4 Dart~    202   136 none       white      yellow          41.9 male
 5 Leia~    150    49 brown      light      brown           19   female
 6 Owen~    178   120 brown, gr~ light      blue            52   male
 7 Beru~    165    75 brown      light      blue            47   female
 8 R5-D4     97    32 <NA>       white, red red             NA   <NA>
 9 Bigg~    183    84 black      light      brown           24   male
10 Obi-~    182    77 auburn, w~ fair       blue-gray       57   male
# ... with 77 more rows, and 9 more variables: homeworld <chr>, species <chr>,
#   films <list>, vehicles <list>, starships <list>, mean.height <dbl>,
#   mean.mass <dbl>, bmi.mean <dbl>, bmi <dbl>
```

```r
# count number of rows in group
num.sp.gend <- starwars %>%
  group_by(species, gender) %>%
  summarize(num = n())

# fraction of mass of each character
fr.mass <- starwars %>%
  group_by(species) %>%
  mutate(pct.mass = mass / sum(mass, na.rm = TRUE)) %>%
  ungroup %>%
  select(name, pct.mass)
```

Joining

```r
bmi <- starwars %>%
  group_by(species) %>%
  summarize(bmi = mean(mass / (height / 100) ^ 2, na.rm = TRUE))

num.tall.characters <- starwars %>%
  filter(height > 150) %>%
  group_by(species) %>%
  summarize(num = n()) %>%
  rename(spp = "species")

num.tall.characters %>%
  left_join(bmi, by = c("spp" = "species"))
```

```
# A tibble: 31 x 3
   spp         num   bmi
   <chr>     <int> <dbl>
 1 Besalisk      1  26.0
 2 Cerean        1  20.9
 3 Chagrian      1 NaN
 4 Clawdite      1  19.5
 5 Droid         2  32.7
 6 Geonosian     1  23.9
 7 Gungan        3  16.8
 8 Human        29  25.5
 9 Hutt          1 443.
```

```
10 Iktotchi      1 NaN
# ... with 21 more rows
```

```
final <- starwars %>%
  group_by(species) %>%
  summarize(bmi = mean(mass / (height / 100) ^ 2, na.rm = TRUE)) %>%
  left_join(
    starwars %>%
      filter(height > 150) %>%
      group_by(species) %>%
      summarize(num = n()) %>%
      rename(spp = "species"),
    by = c("spp" = "species")
  )
```

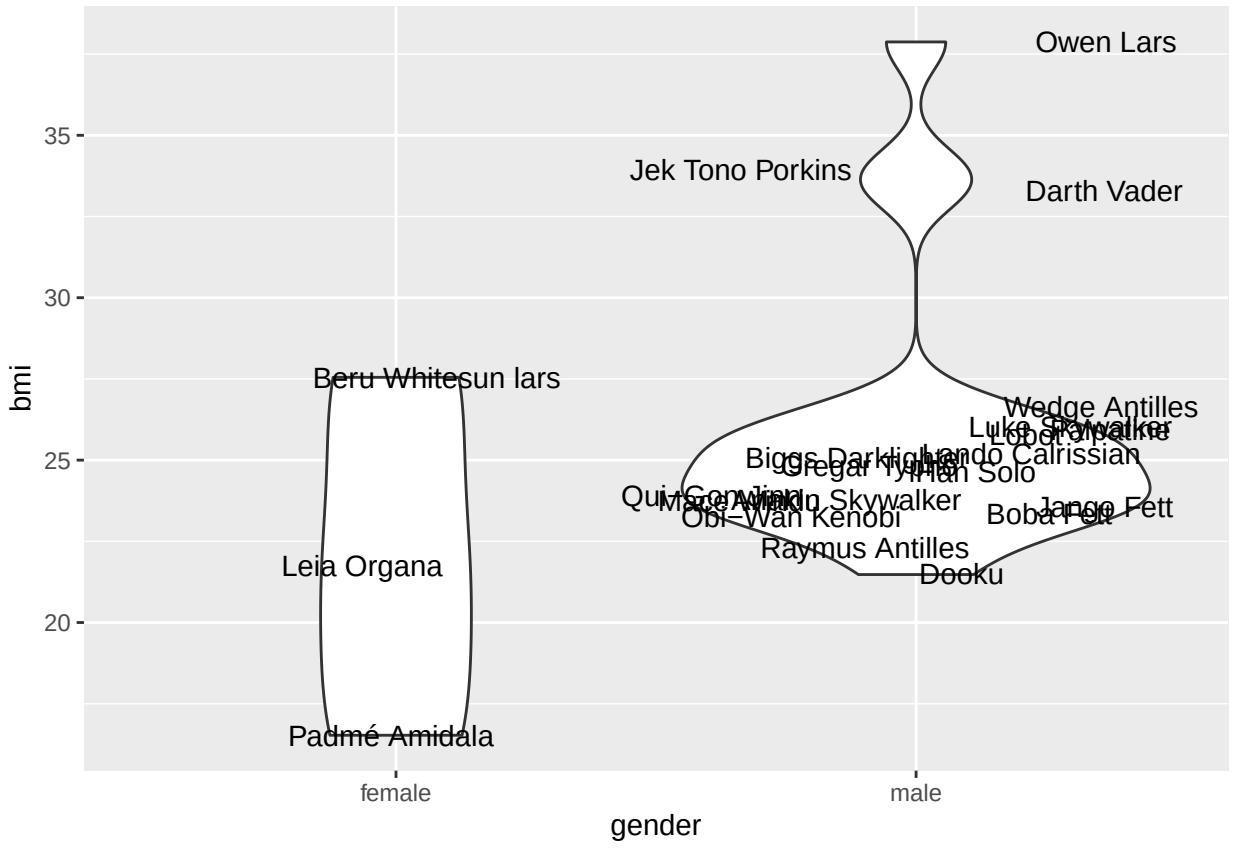
```
Error: `by` can't contain join column `spp` which is missing from LHS
```

tidyr : gather, spread

```
sw <- select(starwars, -(films:starships))

body.colors <- starwars %>%
  select(name, contains("color"))

colors.gathered <- body.colors %>%
  gather(color_type, color, -name) %>%
  arrange(name, color_type, color)

colors.spread <- colors.gathered %>%
  spread(color_type, color) %>%
  as.data.frame
```

pipeline to ggplot

```
starwars %>%
  mutate(bmi = mass / (height / 100) ^ 2) %>%
  select(name, bmi, species, gender) %>%
  filter(complete.cases(.) & species == "Human") %>%
  ggplot(aes(gender, bmi)) +
  geom_violin() +
  geom_text(aes(label = name), position = "jitter")
```


Owen Lars
35
Jek Tono Porkins
Darth Vader
30
Beru Whitesun lars
Wedge Antilles
Luke Skywalker
25
Lando Calrissian
Han Solo
Obi-Wan Kenobi
Boba Fett
Jango Fett
Raymus Antilles
Leia Organa
Dooku
20
Padmé Amidala
female
male
gender
bmi