data_processing

Jonathan Kogan

2022-07-13

Introduction to dplyr

dplyr is an R package in the tidyverse. We can load the package using

```
#install.packages("dplyr") #run this if

# load required libraries
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
?dplyr

# preview dataset
head(starwars)
```

```
## # A tibble: 6 x 14
##   name      height  mass hair_color skin_color eye_color birth_year sex   gender
##   <chr>      <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
## 1 Luke Sky~    172    77 blond      fair       blue            19   male  mascu~
## 2 C-3PO        167    75 <NA>       gold       yellow         112   none  mascu~
## 3 R2-D2         96    32 <NA>       white, bl~ red             33   none  mascu~
## 4 Darth Va~    202   136 none       white      yellow          41.9 male  mascu~
## 5 Leia Org~    150    49 brown      light      brown           19   fema~ femin~
## 6 Owen Lars    178   120 brown, gr~ light      blue            52   male  mascu~
## # ... with 5 more variables: homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

Filtering data

We can filter data using `filter()`. This allows us to subset observations (rows) based on their values (in columns).

Tips:

- Be sure you spell the column name correctly (and the value name if it's a categorical variable). Remember, R is case-sensitive
- Be sure to use `==` when comparing observations. (Remember, `=` is an assignment operator)
- You can use $>$, $<$, $>=$, $<=$ to compare numeric or categorical variables (nominal variables are ranked alphabetically, while ordinal variables have a built-in rank)

```r
# or is |
filter(starwars, hair_color == "blond" | eye_color == "blue")
```

```
## # A tibble: 19 x 14
##    name     height  mass hair_color skin_color eye_color birth_year sex   gender
##    <chr>     <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
##  1 Luke Sk~    172  77   blond      fair       blue            19   male  mascu~
##  2 Owen La~    178 120   brown, gr~ light      blue            52   male  mascu~
##  3 Beru Wh~    165  75   brown      light      blue            47   fema~ femin~
##  4 Anakin ~    188  84   blond      fair       blue            41.9 male  mascu~
##  5 Wilhuff~    180  NA   auburn, g~ fair       blue            64   male  mascu~
##  6 Chewbac~    228 112   brown      unknown    blue           200   male  mascu~
##  7 Jek Ton~    180 110   brown      fair       blue            NA   male  mascu~
##  8 Lobot       175  79   none       light      blue            37   male  mascu~
##  9 Mon Mot~    150  NA   auburn     fair       blue            48   fema~ femin~
## 10 Qui-Gon~    193  89   brown      fair       blue            92   male  mascu~
## 11 Finis V~    170  NA   blond      fair       blue            91   male  mascu~
## 12 Ric Olié    183  NA   brown      fair       blue            NA   <NA>  <NA>
## 13 Adi Gal~    184  50   none       dark       blue            NA   fema~ femin~
## 14 Mas Ame~    196  NA   none       blue       blue            NA   male  mascu~
## 15 Cliegg ~    183  NA   brown      fair       blue            82   male  mascu~
## 16 Luminar~    170  56.2 black      yellow     blue            58   fema~ femin~
## 17 Barriss~    166  50   black      yellow     blue            40   fema~ femin~
## 18 Jocasta~    167  NA   white      fair       blue            NA   fema~ femin~
## 19 Tarfful     234 136   brown      brown      blue            NA   male  mascu~
## # ... with 5 more variables: homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

```r
filter(starwars, hair_color %in% c("blond","blonde"))
```

```
## # A tibble: 4 x 14
##   name      height  mass hair_color skin_color eye_color birth_year sex   gender
##   <chr>      <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
## 1 Luke Sky~    172    77 blond      fair       blue            19   male  mascu~
## 2 Anakin S~    188    84 blond      fair       blue            41.9 male  mascu~
## 3 Finis Va~    170    NA blond      fair       blue            91   male  mascu~
## 4 Zam Wese~    168    55 blonde     fair, gre~ yellow          NA   fema~ femin~
## # ... with 5 more variables: homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

```r
importantPlanets <- c("Tatooine", "Naboo", "Mustafar")
filter(starwars, homeworld %in% importantPlanets)
```

```
## # A tibble: 21 x 14
##    name     height  mass hair_color skin_color eye_color birth_year sex   gender
##    <chr>     <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
##  1 Luke Sk~    172    77 blond      fair       blue            19   male  mascu~
##  2 C-3PO       167    75 <NA>       gold       yellow         112   none  mascu~
##  3 R2-D2        96    32 <NA>       white, bl~ red             33   none  mascu~
##  4 Darth V~    202   136 none       white      yellow          41.9 male  mascu~
##  5 Owen La~    178   120 brown, gr~ light      blue            52   male  mascu~
##  6 Beru Wh~    165    75 brown      light      blue            47   fema~ femin~
##  7 R5-D4        97    32 <NA>       white, red red             NA   none  mascu~
##  8 Biggs D~    183    84 black      light      brown           24   male  mascu~
##  9 Anakin ~    188    84 blond      fair       blue            41.9 male  mascu~
## 10 Palpati~    170    75 grey       pale       yellow          82   male  mascu~
## # ... with 11 more rows, and 5 more variables: homeworld <chr>, species <chr>,
## #   films <list>, vehicles <list>, starships <list>
```

```r
filter(starwars, hair_color == "blond" & species == "Human")
```

```
## # A tibble: 3 x 14
##   name      height  mass hair_color skin_color eye_color birth_year sex   gender
##   <chr>      <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
## 1 Luke Sky~    172    77 blond      fair       blue            19   male  mascu~
## 2 Anakin S~    188    84 blond      fair       blue            41.9 male  mascu~
## 3 Finis Va~    170    NA blond      fair       blue            91   male  mascu~
## # ... with 5 more variables: homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

```r
filter(starwars, species == "Human", homeworld == "Tatooine", skin_color == "fair")
```

```
## # A tibble: 4 x 14
##   name      height  mass hair_color skin_color eye_color birth_year sex   gender
##   <chr>      <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
## 1 Luke Sky~    172    77 blond      fair       blue            19   male  mascu~
## 2 Anakin S~    188    84 blond      fair       blue            41.9 male  mascu~
## 3 Shmi Sky~    163    NA black      fair       brown           72   fema~ femin~
## 4 Cliegg L~    183    NA brown      fair       blue            82   male  mascu~
## # ... with 5 more variables: homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

```r
filter(starwars, mass >= 75, mass <=100, hair_color == "brown", height > 170)
```

```
## # A tibble: 3 x 14
##   name      height  mass hair_color skin_color eye_color birth_year sex   gender
##   <chr>      <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
## 1 Han Solo     180    80 brown      fair       brown              29 male  mascu~
## 2 Qui-Gon ~    193    89 brown      fair       blue               92 male  mascu~
## 3 Raymus A~    188    79 brown      light      brown              NA male  mascu~
## # ... with 5 more variables: homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

```
filter(starwars, mass != 75 | is.na(mass), name < "Mace")
```

```
## # A tibble: 45 x 14
##    name     height  mass hair_color skin_color eye_color birth_year sex   gender
##    <chr>     <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
##  1 Luke Sk~    172    77 blond      fair       blue            19   male  mascu~
##  2 Darth V~    202   136 none       white      yellow          41.9 male  mascu~
##  3 Leia Or~    150    49 brown      light      brown           19   fema~ femin~
##  4 Biggs D~    183    84 black      light      brown           24   male  mascu~
##  5 Anakin ~    188    84 blond      fair       blue            41.9 male  mascu~
##  6 Chewbac~    228   112 brown      unknown    blue           200   male  mascu~
##  7 Han Solo    180    80 brown      fair       brown           29   male  mascu~
##  8 Greedo      173    74 <NA>       green      black           44   male  mascu~
##  9 Jabba D~    175  1358 <NA>       green-tan~ orange         600   herm~ mascu~
## 10 Jek Ton~    180   110 brown      fair       blue            NA   male  mascu~
## # ... with 35 more rows, and 5 more variables: homeworld <chr>, species <chr>,
## #   films <list>, vehicles <list>, starships <list>
```

```
filteredData <- filter(starwars, species == "Human", homeworld == "Tatooine", skin_color == "fair")

# ranked data
head(diamonds) # displays first 6 rows
```

```
## # A tibble: 6 x 10
##   carat cut       color clarity depth table price     x     y     z
##   <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 1  0.23 Ideal     E     SI2      61.5    55   326  3.95  3.98  2.43
## 2  0.21 Premium   E     SI1      59.8    61   326  3.89  3.84  2.31
## 3  0.23 Good      E     VS1      56.9    65   327  4.05  4.07  2.31
## 4  0.29 Premium   I     VS2      62.4    58   334  4.2   4.23  2.63
## 5  0.31 Good      J     SI2      63.3    58   335  4.34  4.35  2.75
## 6  0.24 Very Good J     VVS2     62.8    57   336  3.94  3.96  2.48
```

```
class(diamonds$cut) # gives you the specific type of data
```

```
## [1] "ordered" "factor"
```

```
summary(diamonds$cut) # gives you a count of each category or summary statistics if numeric
```

```
##      Fair      Good Very Good   Premium     Ideal
##      1610      4906     12082     13791     21551
```

```
summary(diamonds$carat)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2000  0.4000  0.7000  0.7979  1.0400  5.0100
```

```
head(diamonds$cut) # displays first 6 values (and levels)
```

```
## [1] Ideal     Premium   Good      Premium   Good      Very Good
## Levels: Fair < Good < Very Good < Premium < Ideal
```

```
filter(diamonds, cut > "Good")
```

```
## # A tibble: 47,424 x 10
##    carat cut       color clarity depth table price     x     y     z
##    <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
##  1  0.23 Ideal     E     SI2      61.5    55   326  3.95  3.98  2.43
##  2  0.21 Premium   E     SI1      59.8    61   326  3.89  3.84  2.31
##  3  0.29 Premium   I     VS2      62.4    58   334  4.2   4.23  2.63
##  4  0.24 Very Good J     VVS2     62.8    57   336  3.94  3.96  2.48
##  5  0.24 Very Good I     VVS1     62.3    57   336  3.95  3.98  2.47
##  6  0.26 Very Good H     SI1      61.9    55   337  4.07  4.11  2.53
##  7  0.23 Very Good H     VS1      59.4    61   338  4     4.05  2.39
##  8  0.23 Ideal     J     VS1      62.8    56   340  3.93  3.9   2.46
##  9  0.22 Premium   F     SI1      60.4    61   342  3.88  3.84  2.33
## 10  0.31 Ideal     J     SI2      62.2    54   344  4.35  4.37  2.71
## # ... with 47,414 more rows
```

```
## Ordering categorical data
unique(starwars$eye_color)
```

```
##  [1] "blue"          "yellow"        "red"           "brown"
##  [5] "blue-gray"     "black"         "orange"        "hazel"
##  [9] "pink"          "unknown"       "red, blue"     "gold"
## [13] "green, yellow" "white"         "dark"
```

```
factor(starwars$eye_color,
       c("red","orange","gold","yellow","green,yellow", "blue", "black"),
       ordered = T)
```

```
##  [1] blue   yellow red    yellow <NA>   blue   blue   red    <NA>   <NA>
## [11] blue   blue   blue   <NA>   black  orange <NA>   blue   <NA>   yellow
## [21] <NA>   red    red    <NA>   blue   orange blue   <NA>   <NA>   black
## [31] blue   red    blue   orange orange orange blue   yellow orange <NA>
## [41] <NA>   yellow <NA>   <NA>   yellow black  orange <NA>   yellow black
## [51] <NA>   blue   orange yellow black  blue   <NA>   <NA>   blue   yellow
## [61] blue   blue   <NA>   <NA>   <NA>   <NA>   yellow yellow black  black
## [71] blue   <NA>   <NA>   <NA>   gold   black  <NA>   blue   <NA>   <NA>
## [81] black  <NA>   <NA>   <NA>   black  <NA>   <NA>
## Levels: red < orange < gold < yellow < green,yellow < blue < black
```

```
### Practice
## Find all characters that are shorter than 100 cm
filter(starwars, height < 100, species != "Droid")
```

```
## # A tibble: 4 x 14
##   name      height  mass hair_color skin_color eye_color birth_year sex   gender
##   <chr>      <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
## 1 Yoda          66    17 white      green      brown            896 male  mascu~
## 2 Wicket S~      88    20 brown      brown      brown              8 male   mascu~
## 3 Dud Bolt       94    45 none       blue, grey yellow             NA male   mascu~
## 4 Ratts Ty~      79    15 none       grey, blue unknown            NA male   mascu~
## # ... with 5 more variables: homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

```
## Which characters were born between the years 100 and 200 (inclusive)?
filter(starwars, birth_year <= 200, birth_year >= 100)
```

```
## # A tibble: 3 x 14
##   name      height  mass hair_color skin_color eye_color birth_year sex   gender
##   <chr>      <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
## 1 C-3PO        167    75 <NA>       gold       yellow           112 none  mascu~
## 2 Chewbacca    228   112 brown      unknown    blue             200 male  mascu~
## 3 Dooku        193    80 white      fair       brown            102 male  mascu~
## # ... with 5 more variables: homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

```
filter(starwars, between(birth_year, 100, 200))
```

```
## # A tibble: 3 x 14
##   name      height  mass hair_color skin_color eye_color birth_year sex   gender
##   <chr>      <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
## 1 C-3PO        167    75 <NA>       gold       yellow           112 none  mascu~
## 2 Chewbacca    228   112 brown      unknown    blue             200 male  mascu~
## 3 Dooku        193    80 white      fair       brown            102 male  mascu~
## # ... with 5 more variables: homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

```
## Which characters weigh over 100kg but are shorter than 185cm?
filter(starwars, mass > 100, height < 185)
```

```
## # A tibble: 3 x 14
##   name      height  mass hair_color skin_color eye_color birth_year sex   gender
##   <chr>      <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
## 1 Owen Lars    178   120 brown, gr~ light      blue              52 male  mascu~
## 2 Jabba De~    175  1358 <NA>       green-tan~ orange           600 herm~ mascu~
## 3 Jek Tono~    180   110 brown      fair       blue              NA male  mascu~
## # ... with 5 more variables: homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

```
## Which characters are missing a hair color?
filter(starwars, is.na(hair_color))
```

```
## # A tibble: 5 x 14
##   name      height  mass hair_color skin_color eye_color birth_year sex   gender
##   <chr>      <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
## 1 C-3PO        167    75 <NA>       gold       yellow           112 none  mascu~
## 2 R2-D2         96    32 <NA>       white, bl~ red               33 none  mascu~
## 3 R5-D4         97    32 <NA>       white, red red               NA none  mascu~
## 4 Greedo       173    74 <NA>       green      black             44 male  mascu~
## 5 Jabba De~    175  1358 <NA>       green-tan~ orange           600 herm~ mascu~
## # ... with 5 more variables: homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

```
mass <- NA
mass ==  10 # returns NA
```

```
## [1] NA
```

```
is.na(mass) # return true
```

```
## [1] TRUE
```

Arranging data

`arrange()` reorders rows. It does not remove any rows. NA values are always at the end when you order by a column.

```
# lowest to highest birth_year
arrange(starwars, birth_year)
```

```
## # A tibble: 87 x 14
##    name     height  mass hair_color skin_color eye_color birth_year sex   gender
##    <chr>     <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
##  1 Wicket ~     88  20   brown      brown      brown            8   male  mascu~
##  2 IG-88       200 140   none       metal      red             15   none  mascu~
##  3 Luke Sk~    172  77   blond      fair       blue            19   male  mascu~
##  4 Leia Or~    150  49   brown      light      brown           19   fema~ femin~
##  5 Wedge A~    170  77   brown      fair       hazel           21   male  mascu~
##  6 Plo Koon    188  80   none       orange     black           22   male  mascu~
##  7 Biggs D~    183  84   black      light      brown           24   male  mascu~
##  8 Han Solo    180  80   brown      fair       brown           29   male  mascu~
##  9 Lando C~    177  79   black      dark       brown           31   male  mascu~
## 10 Boba Fe~    183  78.2 black      fair       brown           31.5 male  mascu~
## # ... with 77 more rows, and 5 more variables: homeworld <chr>, species <chr>,
## #   films <list>, vehicles <list>, starships <list>
```

```
# highest to lowest birth year
arrange(starwars, desc(birth_year))
```

```
## # A tibble: 87 x 14
##    name     height  mass hair_color skin_color eye_color birth_year sex   gender
##    <chr>     <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
##  1 Yoda         66    17 white      green      brown            896 male  mascu~
##  2 Jabba D~    175  1358 <NA>       green-tan~ orange           600 herm~ mascu~
##  3 Chewbac~    228   112 brown      unknown    blue             200 male  mascu~
##  4 C-3PO       167    75 <NA>       gold       yellow           112 none  mascu~
##  5 Dooku       193    80 white      fair       brown            102 male  mascu~
##  6 Qui-Gon~    193    89 brown      fair       blue              92 male  mascu~
##  7 Ki-Adi-~    198    82 white      pale       yellow            92 male  mascu~
##  8 Finis V~    170    NA blond      fair       blue              91 male  mascu~
##  9 Palpati~    170    75 grey       pale       yellow            82 male  mascu~
## 10 Cliegg ~    183    NA brown      fair       blue              82 male  mascu~
## # ... with 77 more rows, and 5 more variables: homeworld <chr>, species <chr>,
## #   films <list>, vehicles <list>, starships <list>
```

```
# categorical is alphabetical
arrange(starwars, hair_color)
```

```
## # A tibble: 87 x 14
##    name     height  mass hair_color skin_color eye_color birth_year sex   gender
##    <chr>     <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
##  1 Mon Mot~    150  NA   auburn     fair       blue            48   fema~ femin~
##  2 Wilhuff~    180  NA   auburn, g~ fair       blue            64   male  mascu~
##  3 Obi-Wan~    182  77   auburn, w~ fair       blue-gray       57   male  mascu~
##  4 Biggs D~    183  84   black      light      brown           24   male  mascu~
##  5 Boba Fe~    183  78.2 black      fair       brown           31.5 male  mascu~
##  6 Lando C~    177  79   black      dark       brown           31   male  mascu~
##  7 Watto       137  NA   black      blue, grey yellow          NA   male  mascu~
##  8 Quarsh ~    183  NA   black      dark       brown           62   <NA>  <NA>
##  9 Shmi Sk~    163  NA   black      fair       brown           72   fema~ femin~
## 10 Eeth Ko~    171  NA   black      brown      brown           NA   male  mascu~
## # ... with 77 more rows, and 5 more variables: homeworld <chr>, species <chr>,
## #   films <list>, vehicles <list>, starships <list>
```

```
# multiple columns
arrange(starwars, hair_color, birth_year)
```

```
## # A tibble: 87 x 14
##    name     height  mass hair_color skin_color eye_color birth_year sex   gender
##    <chr>     <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
##  1 Mon Mot~    150  NA   auburn     fair       blue            48   fema~ femin~
##  2 Wilhuff~    180  NA   auburn, g~ fair       blue            64   male  mascu~
##  3 Obi-Wan~    182  77   auburn, w~ fair       blue-gray       57   male  mascu~
##  4 Biggs D~    183  84   black      light      brown           24   male  mascu~
##  5 Lando C~    177  79   black      dark       brown           31   male  mascu~
##  6 Boba Fe~    183  78.2 black      fair       brown           31.5 male  mascu~
##  7 Barriss~    166  50   black      yellow     blue            40   fema~ femin~
##  8 Luminar~    170  56.2 black      yellow     blue            58   fema~ femin~
##  9 Quarsh ~    183  NA   black      dark       brown           62   <NA>  <NA>
## 10 Jango F~    183  79   black      tan        brown           66   male  mascu~
## # ... with 77 more rows, and 5 more variables: homeworld <chr>, species <chr>,
## #   films <list>, vehicles <list>, starships <list>
```

```
### Practice!

## Arrange starwars characters to find the tallest characters and the shortest characters
arrange(starwars, height)
```

```
## # A tibble: 87 x 14
##    name     height  mass hair_color skin_color eye_color birth_year sex   gender
##    <chr>     <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
##  1 Yoda         66    17 white      green      brown            896 male  mascu~
##  2 Ratts T~     79    15 none       grey, blue unknown           NA male  mascu~
##  3 Wicket ~     88    20 brown      brown      brown              8 male  mascu~
##  4 Dud Bolt     94    45 none       blue, grey yellow            NA male  mascu~
##  5 R2-D2        96    32 <NA>       white, bl~ red               33 none  mascu~
##  6 R4-P17       96    NA none       silver, r~ red, blue         NA none  femin~
##  7 R5-D4        97    32 <NA>       white, red red               NA none  mascu~
##  8 Sebulba     112    40 none       grey, red  orange            NA male  mascu~
##  9 Gasgano     122    NA none       white, bl~ black             NA male  mascu~
## 10 Watto       137    NA black      blue, grey yellow            NA male  mascu~
## # ... with 77 more rows, and 5 more variables: homeworld <chr>, species <chr>,
## #   films <list>, vehicles <list>, starships <list>
```

```
arrange(starwars, desc(height))
```

```
## # A tibble: 87 x 14
##    name     height  mass hair_color skin_color eye_color birth_year sex   gender
##    <chr>     <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
##  1 Yarael ~    264    NA none       white      yellow          NA   male  mascu~
##  2 Tarfful     234   136 brown      brown      blue            NA   male  mascu~
##  3 Lama Su     229    88 none       grey       black           NA   male  mascu~
##  4 Chewbac~    228   112 brown      unknown    blue           200   male  mascu~
##  5 Roos Ta~    224    82 none       grey       orange          NA   male  mascu~
##  6 Grievous    216   159 none       brown, wh~ green, y~       NA   male  mascu~
##  7 Taun We     213    NA none       grey       black           NA   fema~ femin~
##  8 Rugor N~    206    NA none       green      orange          NA   male  mascu~
##  9 Tion Me~    206    80 none       grey       black           NA   male  mascu~
## 10 Darth V~    202   136 none       white      yellow          41.9 male  mascu~
## # ... with 77 more rows, and 5 more variables: homeworld <chr>, species <chr>,
## #   films <list>, vehicles <list>, starships <list>
```

```
## Alphabetize the star wars characters by name
arrange(starwars, name)
```

```
## # A tibble: 87 x 14
##    name     height  mass hair_color skin_color eye_color birth_year sex   gender
##    <chr>     <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
##  1 Ackbar      180    83 none       brown mot~ orange          41   male  mascu~
##  2 Adi Gal~    184    50 none       dark       blue            NA   fema~ femin~
##  3 Anakin ~    188    84 blond      fair       blue            41.9 male  mascu~
##  4 Arvel C~     NA    NA brown      fair       brown           NA   male  mascu~
##  5 Ayla Se~    178    55 none       blue       hazel           48   fema~ femin~
##  6 Bail Pr~    191    NA black      tan        brown           67   male  mascu~
##  7 Barriss~    166    50 black      yellow     blue            40   fema~ femin~
##  8 BB8          NA    NA none       none       black           NA   none  mascu~
##  9 Ben Qua~    163    65 none       grey, gre~ orange          NA   male  mascu~
## 10 Beru Wh~    165    75 brown      light      blue            47   fema~ femin~
## # ... with 77 more rows, and 5 more variables: homeworld <chr>, species <chr>,
## #   films <list>, vehicles <list>, starships <list>
```

```
## How could you use arrange() to sort all missing values to the start?
arrange(starwars, desc(is.na(height)), desc(is.na(mass)))
```

```
## # A tibble: 87 x 14
##    name     height  mass hair_color skin_color eye_color birth_year sex   gender
##    <chr>     <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
##  1 Arvel C~     NA    NA brown      fair       brown             NA male  mascu~
##  2 Finn         NA    NA black      dark       dark              NA male  mascu~
##  3 Rey          NA    NA brown      light      hazel             NA fema~ femin~
##  4 Poe Dam~     NA    NA brown      light      brown             NA male  mascu~
##  5 BB8          NA    NA none       none       black             NA none  mascu~
##  6 Captain~     NA    NA unknown    unknown    unknown           NA <NA>  <NA>
##  7 Wilhuff~    180    NA auburn, g~ fair       blue              64 male  mascu~
##  8 Mon Mot~    150    NA auburn     fair       blue              48 fema~ femin~
##  9 Finis V~    170    NA blond      fair       blue              91 male  mascu~
## 10 Rugor N~    206    NA none       green      orange            NA male  mascu~
## # ... with 77 more rows, and 5 more variables: homeworld <chr>, species <chr>,
## #   films <list>, vehicles <list>, starships <list>
```

```
arrange(starwars, desc(is.na(starwars)))
```

```
## # A tibble: 87 x 14
##    name     height  mass hair_color skin_color eye_color birth_year sex   gender
##    <chr>     <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
##  1 Captain~     NA    NA unknown    unknown    unknown           NA <NA>  <NA>
##  2 Arvel C~     NA    NA brown      fair       brown             NA male  mascu~
##  3 Finn         NA    NA black      dark       dark              NA male  mascu~
##  4 Rey          NA    NA brown      light      hazel             NA fema~ femin~
##  5 Poe Dam~     NA    NA brown      light      brown             NA male  mascu~
##  6 BB8          NA    NA none       none       black             NA none  mascu~
##  7 Ric Olié    183    NA brown      fair       blue              NA <NA>  <NA>
##  8 R4-P17       96    NA none       silver, r~ red, blue         NA none  femin~
##  9 Rugor N~    206    NA none       green      orange            NA male  mascu~
## 10 Watto       137    NA black      blue, grey yellow            NA male  mascu~
## # ... with 77 more rows, and 5 more variables: homeworld <chr>, species <chr>,
## #   films <list>, vehicles <list>, starships <list>
```

We can select certain columns in the dataset

`select()` allows us to retain only certain variables (columns). It doesn't change the order, but it removes columns not named

```
select(starwars, hair_color, skin_color, eye_color)
```

```
## # A tibble: 87 x 3
##    hair_color    skin_color  eye_color
##    <chr>         <chr>       <chr>
##  1 blond         fair        blue
##  2 <NA>          gold        yellow
##  3 <NA>          white, blue red
##  4 none          white       yellow
##  5 brown         light       brown
##  6 brown, grey   light       blue
##  7 brown         light       blue
##  8 <NA>          white, red  red
##  9 black         light       brown
## 10 auburn, white fair        blue-gray
## # ... with 77 more rows
```

```
head(starwars)
```

```
## # A tibble: 6 x 14
##   name      height  mass hair_color skin_color eye_color birth_year sex   gender
##   <chr>      <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
## 1 Luke Sky~    172    77 blond      fair       blue            19   male  mascu~
## 2 C-3PO        167    75 <NA>       gold       yellow         112   none  mascu~
## 3 R2-D2         96    32 <NA>       white, bl~ red             33   none  mascu~
## 4 Darth Va~    202   136 none       white      yellow          41.9 male  mascu~
## 5 Leia Org~    150    49 brown      light      brown           19   fema~ femin~
## 6 Owen Lars    178   120 brown, gr~ light      blue            52   male  mascu~
## # ... with 5 more variables: homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

```
select(starwars, hair_color:eye_color) # returns every column between first:last
```

```
## # A tibble: 87 x 3
##    hair_color    skin_color  eye_color
##    <chr>         <chr>       <chr>
##  1 blond         fair        blue
##  2 <NA>          gold        yellow
##  3 <NA>          white, blue red
##  4 none          white       yellow
##  5 brown         light       brown
##  6 brown, grey   light       blue
##  7 brown         light       blue
##  8 <NA>          white, red  red
##  9 black         light       brown
## 10 auburn, white fair        blue-gray
## # ... with 77 more rows
```

```
select(starwars, -hair_color)
```

```
## # A tibble: 87 x 13
##    name      height  mass skin_color eye_color birth_year sex   gender homeworld
##    <chr>      <int> <dbl> <chr>      <chr>          <dbl> <chr> <chr>  <chr>
##  1 Luke Sky~    172    77 fair       blue            19   male  mascu~ Tatooine
##  2 C-3PO        167    75 gold       yellow         112   none  mascu~ Tatooine
##  3 R2-D2         96    32 white, bl~ red             33   none  mascu~ Naboo
##  4 Darth Va~    202   136 white      yellow          41.9 male  mascu~ Tatooine
##  5 Leia Org~    150    49 light      brown           19   fema~ femin~ Alderaan
##  6 Owen Lars    178   120 light      blue            52   male  mascu~ Tatooine
##  7 Beru Whi~    165    75 light      blue            47   fema~ femin~ Tatooine
##  8 R5-D4         97    32 white, red red             NA   none  mascu~ Tatooine
##  9 Biggs Da~    183    84 light      brown           24   male  mascu~ Tatooine
## 10 Obi-Wan ~    182    77 fair       blue-gray       57   male  mascu~ Stewjon
## # ... with 77 more rows, and 4 more variables: species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

```
select(starwars, -(hair_color:eye_color))
```

```
## # A tibble: 87 x 11
##    name    height  mass birth_year sex   gender homeworld species films vehicles
##    <chr>    <int> <dbl>      <dbl> <chr> <chr>  <chr>     <chr>   <lis> <list>
##  1 Luke S~    172    77       19   male  mascu~ Tatooine  Human   <chr> <chr>
##  2 C-3PO      167    75      112   none  mascu~ Tatooine  Droid   <chr> <chr>
##  3 R2-D2       96    32       33   none  mascu~ Naboo     Droid   <chr> <chr>
##  4 Darth ~    202   136       41.9 male  mascu~ Tatooine  Human   <chr> <chr>
##  5 Leia O~    150    49       19   fema~ femin~ Alderaan  Human   <chr> <chr>
##  6 Owen L~    178   120       52   male  mascu~ Tatooine  Human   <chr> <chr>
##  7 Beru W~    165    75       47   fema~ femin~ Tatooine  Human   <chr> <chr>
##  8 R5-D4       97    32       NA   none  mascu~ Tatooine  Droid   <chr> <chr>
##  9 Biggs ~    183    84       24   male  mascu~ Tatooine  Human   <chr> <chr>
## 10 Obi-Wa~    182    77       57   male  mascu~ Stewjon   Human   <chr> <chr>
## # ... with 77 more rows, and 1 more variable: starships <list>
```

```
starwars_no_color <- select(starwars, -(hair_color:eye_color))
#ggplot(starwars_no_color, aes(x = hair_color)) # error because we removed it

select(starwars, contains("color"))
```

```
## # A tibble: 87 x 3
##    hair_color    skin_color  eye_color
##    <chr>         <chr>       <chr>
##  1 blond         fair        blue
##  2 <NA>          gold        yellow
##  3 <NA>          white, blue red
##  4 none          white       yellow
##  5 brown         light       brown
##  6 brown, grey   light       blue
##  7 brown         light       blue
##  8 <NA>          white, red  red
##  9 black         light       brown
## 10 auburn, white fair        blue-gray
## # ... with 77 more rows
```

```
select(starwars, ends_with("color"))
```

```
## # A tibble: 87 x 3
##    hair_color    skin_color  eye_color
##    <chr>         <chr>       <chr>
##  1 blond         fair        blue
##  2 <NA>          gold        yellow
##  3 <NA>          white, blue red
##  4 none          white       yellow
##  5 brown         light       brown
##  6 brown, grey   light       blue
##  7 brown         light       blue
##  8 <NA>          white, red  red
##  9 black         light       brown
## 10 auburn, white fair        blue-gray
## # ... with 77 more rows
```

```
select(starwars, contains("_"))
```

```
## # A tibble: 87 x 4
##    hair_color    skin_color  eye_color birth_year
##    <chr>         <chr>       <chr>          <dbl>
##  1 blond         fair        blue            19
##  2 <NA>          gold        yellow         112
##  3 <NA>          white, blue red             33
##  4 none          white       yellow          41.9
##  5 brown         light       brown           19
##  6 brown, grey   light       blue            52
##  7 brown         light       blue            47
##  8 <NA>          white, red  red             NA
##  9 black         light       brown           24
## 10 auburn, white fair        blue-gray       57
## # ... with 77 more rows
```

```
select(starwars, starts_with("s"), ends_with("color"))
```

```
## # A tibble: 87 x 6
##    skin_color  sex    species starships hair_color    eye_color
##    <chr>       <chr>  <chr>   <list>    <chr>         <chr>
##  1 fair        male   Human   <chr [2]> blond         blue
##  2 gold        none   Droid   <chr [0]> <NA>          yellow
##  3 white, blue none   Droid   <chr [0]> <NA>          red
##  4 white       male   Human   <chr [1]> none          yellow
##  5 light       female Human   <chr [0]> brown         brown
##  6 light       male   Human   <chr [0]> brown, grey   blue
##  7 light       female Human   <chr [0]> brown         blue
##  8 white, red  none   Droid   <chr [0]> <NA>          red
##  9 light       male   Human   <chr [1]> black         brown
## 10 fair        male   Human   <chr [5]> auburn, white blue-gray
## # ... with 77 more rows
```

```
?select

starwars2 <- rename(starwars, birthYear = birth_year)
starwars2
```

```
## # A tibble: 87 x 14
##    name      height  mass hair_color skin_color eye_color birthYear sex   gender
##    <chr>      <int> <dbl> <chr>      <chr>      <chr>         <dbl> <chr> <chr>
##  1 Luke Sky~    172    77 blond      fair       blue           19   male  mascu~
##  2 C-3PO        167    75 <NA>       gold       yellow        112   none  mascu~
##  3 R2-D2         96    32 <NA>       white, bl~ red            33   none  mascu~
##  4 Darth Va~    202   136 none       white      yellow         41.9 male  mascu~
##  5 Leia Org~    150    49 brown      light      brown          19   fema~ femin~
##  6 Owen Lars    178   120 brown, gr~ light      blue           52   male  mascu~
##  7 Beru Whi~    165    75 brown      light      blue           47   fema~ femin~
##  8 R5-D4         97    32 <NA>       white, red red            NA   none  mascu~
##  9 Biggs Da~    183    84 black      light      brown          24   male  mascu~
## 10 Obi-Wan ~    182    77 auburn, w~ fair       blue-gray      57   male  mascu~
## # ... with 77 more rows, and 5 more variables: homeworld <chr>, species <chr>,
## #   films <list>, vehicles <list>, starships <list>
```

```
### Practice!!

## Select out the homeworld and species for the starwars dataset. What question might this subset of va
select(starwars, homeworld, species)
```

```
## # A tibble: 87 x 2
##    homeworld species
##    <chr>     <chr>  
##  1 Tatooine  Human  
##  2 Tatooine  Droid  
##  3 Naboo     Droid  
##  4 Tatooine  Human  
##  5 Alderaan  Human  
##  6 Tatooine  Human  
##  7 Tatooine  Human  
##  8 Tatooine  Droid  
##  9 Tatooine  Human  
## 10 Stewjon   Human  
## # ... with 77 more rows
```

```
## Brainstorm as many ways as possible to select hair_color, eye_color, skin_color, and birth_year
select(starwars, ends_with("color"), ends_with("year"))
```

```
## # A tibble: 87 x 4
##    hair_color    skin_color  eye_color birth_year
##    <chr>         <chr>       <chr>          <dbl>
##  1 blond         fair        blue            19  
##  2 <NA>          gold        yellow         112  
##  3 <NA>          white, blue red             33  
##  4 none          white       yellow          41.9
##  5 brown         light       brown           19  
##  6 brown, grey   light       blue            52  
##  7 brown         light       blue            47  
##  8 <NA>          white, red  red             NA  
##  9 black         light       brown           24  
## 10 auburn, white fair        blue-gray       57  
## # ... with 77 more rows
```

Adding new columns

`mutate()` adds new columns to the end of your dataset.

```
starwars_small <- select(starwars, height, mass, birth_year)
head(starwars_small)
```

```
## # A tibble: 6 x 3
##   height  mass birth_year
##    <int> <dbl>      <dbl>
## 1    172    77       19
## 2    167    75      112
## 3     96    32       33
## 4    202   136       41.9
## 5    150    49       19
## 6    178   120       52
```

```
starwars_small <- mutate(starwars_small, height_m = height/100)
mutate(starwars_small, bmi = mass/(height_m^2))
```

```
## # A tibble: 87 x 5
##    height  mass birth_year height_m   bmi
##     <int> <dbl>      <dbl>    <dbl> <dbl>
##  1    172    77       19       1.72  26.0
##  2    167    75      112       1.67  26.9
##  3     96    32       33       0.96  34.7
##  4    202   136       41.9     2.02  33.3
##  5    150    49       19       1.5   21.8
##  6    178   120       52       1.78  37.9
##  7    165    75       47       1.65  27.5
##  8     97    32       NA       0.97  34.0
##  9    183    84       24       1.83  25.1
## 10    182    77       57       1.82  23.2
## # ... with 77 more rows
```

```
starwars_small <- select(starwars, height, mass, birth_year)
mutate(starwars_small,
       height_m = height/100,
       bmi = mass/(height_m^2))
```

```
## # A tibble: 87 x 5
##    height  mass birth_year height_m   bmi
##     <int> <dbl>      <dbl>    <dbl> <dbl>
##  1    172    77       19       1.72  26.0
##  2    167    75      112       1.67  26.9
##  3     96    32       33       0.96  34.7
##  4    202   136       41.9     2.02  33.3
##  5    150    49       19       1.5   21.8
##  6    178   120       52       1.78  37.9
##  7    165    75       47       1.65  27.5
##  8     97    32       NA       0.97  34.0
##  9    183    84       24       1.83  25.1
## 10    182    77       57       1.82  23.2
## # ... with 77 more rows
```

```
# to only keep new columns, use transmute
transmute(starwars_small,
       height_m = height/100,
       bmi = mass/(height_m^2))
```

```
## # 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
```

```
## using aggregate functions
prop_mass <- mutate(starwars_small, proportional_mass = mass/sum(mass, na.rm = T))
arrange(prop_mass, desc(proportional_mass))
```

```
## # A tibble: 87 x 4
##    height  mass birth_year proportional_mass
##     <int> <dbl>      <dbl>             <dbl>
##  1    175  1358      600              0.237
##  2    216   159       NA              0.0277
##  3    200   140       15              0.0244
##  4    202   136       41.9            0.0237
##  5    234   136       NA              0.0237
##  6    178   120       52              0.0209
##  7    190   113       53              0.0197
##  8    228   112      200              0.0195
##  9    180   110       NA              0.0192
## 10    198   102       NA              0.0178
## # ... with 77 more rows
```

###Summarizing and grouping dara

`summarize()` collapses an entire column of data to a single value

```
mutate(starwars, mean_mass = mean(mass, na.rm = T))
```

```
## # A tibble: 87 x 15
##    name     height  mass hair_color skin_color eye_color birth_year sex   gender
##    <chr>     <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
##  1 Luke Sk~    172    77 blond      fair       blue            19   male  mascu~
##  2 C-3PO       167    75 <NA>       gold       yellow         112   none  mascu~
##  3 R2-D2        96    32 <NA>       white, bl~ red             33   none  mascu~
##  4 Darth V~    202   136 none       white      yellow          41.9 male  mascu~
##  5 Leia Or~    150    49 brown      light      brown           19   fema~ femin~
##  6 Owen La~    178   120 brown, gr~ light      blue            52   male  mascu~
##  7 Beru Wh~    165    75 brown      light      blue            47   fema~ femin~
##  8 R5-D4        97    32 <NA>       white, red red             NA   none  mascu~
##  9 Biggs D~    183    84 black      light      brown           24   male  mascu~
## 10 Obi-Wan~    182    77 auburn, w~ fair       blue-gray       57   male  mascu~
## # ... with 77 more rows, and 6 more variables: homeworld <chr>, species <chr>,
## #   films <list>, vehicles <list>, starships <list>, mean_mass <dbl>
```

```
summarise(starwars, mean_mass = mean(mass, na.rm = T))
```

```
## # A tibble: 1 x 1
##   mean_mass
##       <dbl>
## 1      97.3
```

```
mean(starwars$mass, na.rm = T)
```

```
## [1] 97.31186
```

```
species_masses <- summarise(group_by(starwars, species), mean_mass = mean(mass, na.rm = T), count = n()

# the pipe operator %>%
# function(x, y) is the same as x %>% function(y)
# When using dplyr functions, generally always start with the dataset
species_masses <- starwars %>%
  group_by(species) %>%
  summarise(mean_mass = mean(mass, na.rm = T),
            count = n()) %>%
  arrange(desc(mean_mass))

species_masses
```

```
## # A tibble: 38 x 3
##    species      mean_mass count
##    <chr>            <dbl> <int>
##  1 Hutt            1358       1
##  2 Kaleesh          159       1
##  3 Wookiee          124       2
##  4 Trandoshan       113       1
##  5 Besalisk         102       1
##  6 Neimodian         90       1
##  7 Kaminoan          88       2
##  8 Nautolan          87       1
##  9 Mon Calamari      83       1
## 10 Human             82.8    35
## # ... with 28 more rows
```

```
arrange(species_masses, desc(mean_mass))
```

```
## # A tibble: 38 x 3
##    species      mean_mass count
##    <chr>            <dbl> <int>
##  1 Hutt            1358       1
##  2 Kaleesh          159       1
##  3 Wookiee          124       2
##  4 Trandoshan       113       1
##  5 Besalisk         102       1
##  6 Neimodian         90       1
##  7 Kaminoan          88       2
##  8 Nautolan          87       1
##  9 Mon Calamari      83       1
## 10 Human             82.8    35
## # ... with 28 more rows
```

Sampling a designated number of rows

`sample_n()` allows us to sample a random number of rows from our dataset.

```
# 10 random rows
starwars_10rows <- sample_n(starwars, 10)

starwars_10rows
```

```
## # A tibble: 10 x 14
##    name     height  mass hair_color skin_color eye_color birth_year sex   gender
##    <chr>     <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
##  1 Arvel C~     NA    NA brown      fair       brown             NA male  mascu~
##  2 Zam Wes~    168    55 blonde     fair, gre~ yellow            NA fema~ femin~
##  3 Nien Nu~    160    68 none       grey       black             NA male  mascu~
##  4 Ki-Adi-~    198    82 white      pale       yellow            92 male  mascu~
##  5 Mas Ame~    196    NA none       blue       blue              NA male  mascu~
##  6 C-3PO       167    75 <NA>       gold       yellow           112 none  mascu~
##  7 R5-D4        97    32 <NA>       white, red red               NA none  mascu~
##  8 Padmé A~    165    45 brown      light      brown             46 fema~ femin~
##  9 Cordé       157    NA brown      light      brown             NA fema~ femin~
## 10 Finn         NA    NA black      dark       dark              NA male  mascu~
## # ... with 5 more variables: homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

```
# 10% of rows, randomly selected
starwars_10percent <- sample_frac(starwars, 0.1)
starwars_10percent # 9 rows is 10%
```

```
## # A tibble: 9 x 14
##   name      height  mass hair_color skin_color eye_color birth_year sex   gender
##   <chr>      <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
## 1 BB8           NA    NA none       none       black             NA none  mascu~
## 2 Poe Dame~     NA    NA brown      light      brown             NA male  mascu~
## 3 Lando Ca~    177    79 black      dark       brown             31 male  mascu~
## 4 Tion Med~    206    80 none       grey       black             NA male  mascu~
## 5 Rey           NA    NA brown      light      hazel             NA fema~ femin~
## 6 Lama Su      229    88 none       grey       black             NA male  mascu~
## 7 Finis Va~    170    NA blond      fair       blue              91 male  mascu~
## 8 Beru Whi~    165    75 brown      light      blue              47 fema~ femin~
## 9 Captain ~     NA    NA unknown    unknown    unknown           NA <NA>  <NA>
## # ... with 5 more variables: homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

We can also take a "slice" of our dataset using `slice()` and its related set of functions