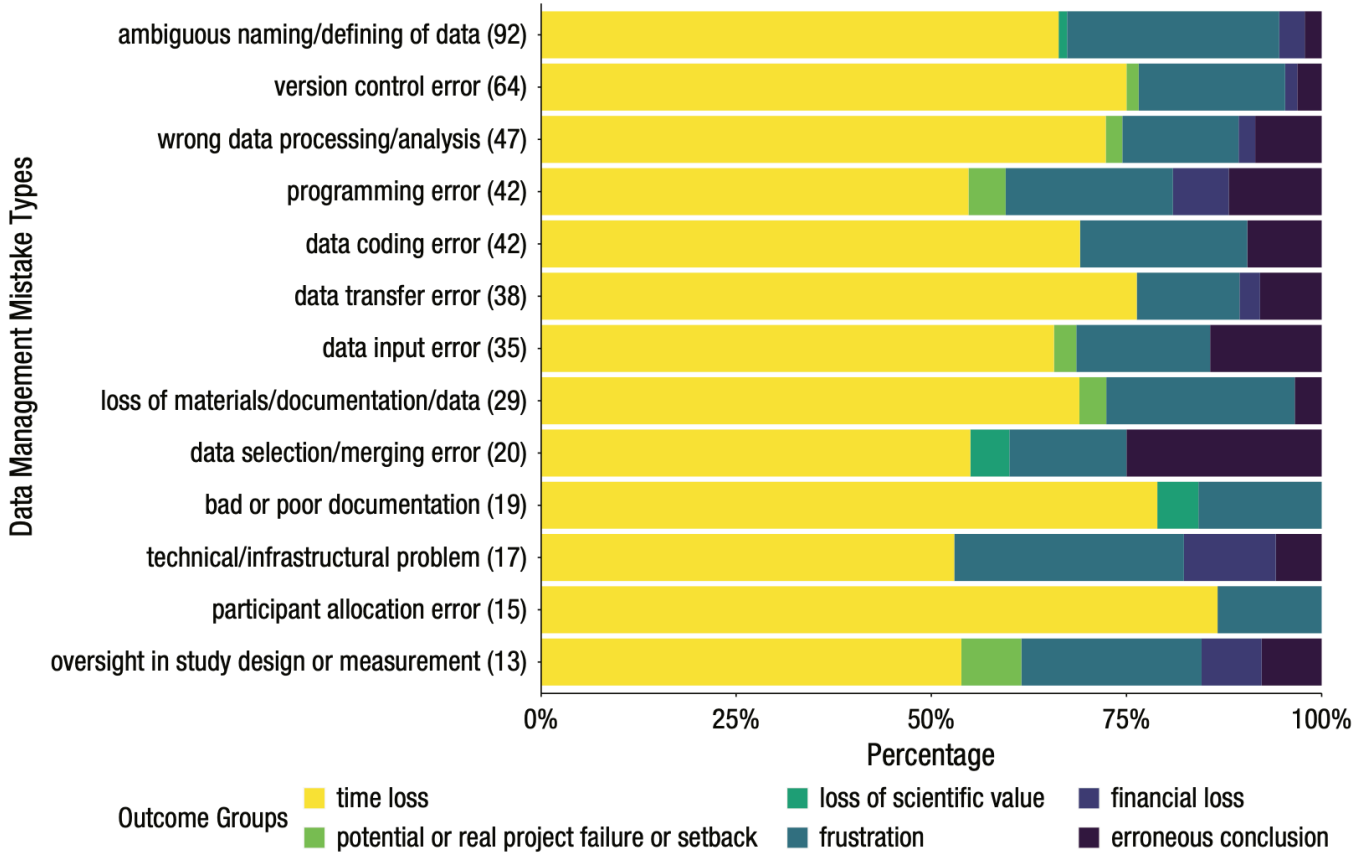
R Data Types and Transformations with dplyr

PSYC 259: Principles of Data Science

John Franchak

How can R data processing help us minimize mistakes?



How can R data processing help us minimize mistakes?

- Code *is* documentation
- Transform data with less human intervention
- But coding errors are also common...what can we do to make sure that our code works?
  - Instead of “hard” coding based on position (such as ds[1, 2]), filter and select by name and logical conditions
  - Use good coding practices to clean up variable names to make them human-readable
  - Reduce code duplication (copy-paste-tweak) with loops and functions
  - Write code that's reproducible (avoiding absolute file paths)

Data type and transformation tutorial

- Data types
- Logical statements
- Introduction to *dplyr*

Follow along from the Github repo

Last updated: 2022-01-20

Data types

- Why have pre-defined types?
  - Allows software to efficiently store data in memory
  - If a value is an integer (1, 2, 3, 4) storing it as an integer makes calculations easier compared to storing it as a double (2.34542480424624086)
- Allows software to implement rules about transformations
  - Addition/subtraction for a date follows different rules compared with integers/double
  - “Less than” makes sense when comparing numbers, but not when comparing strings

Common data types in R

- Numeric
  - integer: 1, 2, 3
  - double: 1.12124
- Character: “hello”
- Logical: T/F (TRUE/FALSE)
- Datetime
- Factor
- Use `typeof()` function to check type of a value, `str()` or `glimpse()` to check the types of each column of a tibble

Functions to test and convert types

- Take a value and convert it into a different type
  - `as.numeric()`, `as.character()`, `as.date()`
- Check if something is a particular type
  - `is.numeric()`, `is.character()`

Data types

```
x <- 1
```

Data types

```
x <- 1
x
```

```
[1] 1
```

Data types

```
x <- 1
x
typeof(x)
```

```
[1] 1

[1] "double"
```

Data types

```
x <- 1
x
typeof(x)
is.numeric(x)
```

```
[1] 1

[1] "double"

[1] TRUE
```

Data types

```
x <- 1
x
typeof(x)
is.numeric(x)
is.character(x)
```

```
[1] 1

[1] "double"

[1] TRUE

[1] FALSE
```

Data types

```
x <- 1
x
typeof(x)
is.numeric(x)
is.character(x)
as.character(x)
```

```
[1] 1

[1] "double"

[1] TRUE

[1] FALSE

[1] "1"
```

Data types

```
x <- 1
x
typeof(x)
is.numeric(x)
is.character(x)
as.character(x)
x + 1
```

```
[1] 1

[1] "double"

[1] TRUE

[1] FALSE

[1] "1"

[1] 2
```

Data types

```
x <- 1
x
typeof(x)
is.numeric(x)
is.character(x)
as.character(x)
x + 1

x <- "data.csv"
```

```
[1] 1

[1] "double"

[1] TRUE

[1] FALSE

[1] "1"

[1] 2
```

Data types

```
x <- 1
x
typeof(x)
is.numeric(x)
is.character(x)
as.character(x)
x + 1

x <- "data.csv"
typeof(x)
```

```
[1] 1

[1] "double"

[1] TRUE

[1] FALSE

[1] "1"

[1] 2

[1] "character"
```

Data types

```
x <- 1
x
typeof(x)
is.numeric(x)
is.character(x)
as.character(x)
x + 1

x <- "data.csv"
typeof(x)
is.numeric(x)
```

```
[1] 1

[1] "double"

[1] TRUE

[1] FALSE

[1] "1"

[1] 2

[1] "character"

[1] FALSE
```

Data types

```
x <- 1
x
typeof(x)
is.numeric(x)
is.character(x)
as.character(x)
x + 1

x <- "data.csv"
typeof(x)
is.numeric(x)
is.character(x)
```

```
[1] 1

[1] "double"

[1] TRUE

[1] FALSE

[1] "1"

[1] 2

[1] "character"

[1] FALSE

[1] TRUE
```

Data types

```
x <- 1
x
typeof(x)
is.numeric(x)
is.character(x)
as.character(x)
x + 1

x <- "data.csv"
typeof(x)
is.numeric(x)
is.character(x)
as.numeric(x)
```

```
[1] 1

[1] "double"

[1] TRUE

[1] FALSE

[1] "1"

[1] 2

[1] "character"

[1] FALSE

[1] TRUE

[1] NA
```

Data types

```r
x <- 1
x
typeof(x)
is.numeric(x)
is.character(x)
as.character(x)
x + 1

x <- "data.csv"
typeof(x)
is.numeric(x)
is.character(x)
as.numeric(x)
#"data_raw" + x  #"data_raw" + x
```

```
[1] 1

[1] "double"

[1] TRUE

[1] FALSE

[1] "1"

[1] 2

[1] "character"

[1] FALSE

[1] TRUE

[1] NA
```

Data types

```
x <- 1
x
typeof(x)
is.numeric(x)
is.character(x)
as.character(x)
x + 1

x <- "data.csv"
typeof(x)
is.numeric(x)
is.character(x)
as.numeric(x)
#"data_raw" + x  #"data_raw" + x
paste0("data_raw/",x)
```

```
[1] 1

[1] "double"

[1] TRUE

[1] FALSE

[1] "1"

[1] 2

[1] "character"

[1] FALSE

[1] TRUE

[1] NA

[1] "data_raw/data.csv"
```

Missing values

- `NA` means missing data in R
- `is.na()` checks whether a value is missing
- not to be confused with `NULL` (empty), or `NaN` (not a number)

```r
x <-  c(1, 2, 3, NA)
```

```r
x <-  c(1, 2, 3, NA)
print(x)
```

```
[1]  1  2  3 NA
```

```r
x <-  c(1, 2, 3, NA)
print(x)
is.na(x)
```

```
[1]  1  2  3 NA

[1] FALSE FALSE FALSE  TRUE
```

```r
x <-  c(1, 2, 3, NA)
print(x)
is.na(x)
mean(x)
```

```
[1]  1  2  3 NA

[1] FALSE FALSE FALSE  TRUE

[1] NA
```

```r
x <-  c(1, 2, 3, NA)
print(x)
is.na(x)
mean(x)
mean(x, na.rm = TRUE)
```

```
[1]  1  2  3 NA

[1] FALSE FALSE FALSE  TRUE

[1] NA

[1] 2
```

```r
x <-  c(1, 2, 3, NA)
print(x)
is.na(x)
mean(x)
mean(x, na.rm = TRUE)

x <- c(1, 2, 3, NULL)
```

```
[1]  1  2  3 NA

[1] FALSE FALSE FALSE  TRUE

[1] NA

[1] 2
```

```r
x <-  c(1, 2, 3, NA)
print(x)
is.na(x)
mean(x)
mean(x, na.rm = TRUE)

x <- c(1, 2, 3, NULL)
x
```

```
[1]  1  2  3 NA

[1] FALSE FALSE FALSE  TRUE

[1] NA

[1] 2

[1] 1 2 3
```

Logical comparisons

- Comparisons between values that result in TRUE/FALSE
- Greater than/less than (>, >=, <, <=)
- Equals (==)
- Not equals (!=)
- Not (!)
- And (&)
- Or (|)
- %in%

Logical comparisons

```
x <- 1
```

Logical comparisons

```
x <- 1
x > 0
```

```
[1] TRUE
```

Logical comparisons

```
x <- 1
x > 0
x > 2
```

```
[1] TRUE

[1] FALSE
```

Logical comparisons

```
x <- 1
x > 0
x > 2
x == 1
```

```
[1] TRUE

[1] FALSE

[1] TRUE
```

Logical comparisons

```
x <- 1
x > 0
x > 2
x == 1
x != 2
```

```
[1] TRUE

[1] FALSE

[1] TRUE

[1] TRUE
```

Logical comparisons

```
x <- 1
x > 0
x > 2
x == 1
x != 2
!(x == x)
```

```
[1] TRUE

[1] FALSE

[1] TRUE

[1] TRUE

[1] FALSE
```

Logical comparisons

```
x <- 1
x > 0
x > 2
x == 1
x != 2
!(x == x)
"s" == "S"
```

```
[1] TRUE

[1] FALSE

[1] TRUE

[1] TRUE

[1] FALSE

[1] FALSE
```

Logical comparisons

```
x <- 1
x > 0
x > 2
x == 1
x != 2
!(x == x)
"s" == "S"
1 > 0 | 0 > 1
```

```
[1] TRUE

[1] FALSE

[1] TRUE

[1] TRUE

[1] FALSE

[1] FALSE

[1] TRUE
```

Logical comparisons

```
x <- 1
x > 0
x > 2
x == 1
x != 2
!(x == x)
"s" == "S"
1 > 0 | 0 > 1
1 > 0 & 0 > 1
```

```
[1] TRUE

[1] FALSE

[1] TRUE

[1] TRUE

[1] FALSE

[1] FALSE

[1] TRUE

[1] FALSE
```

Logical comparisons

```
x <- 1
x > 0
x > 2
x == 1
x != 2
!(x == x)
"s" == "S"
1 > 0 | 0 > 1
1 > 0 & 0 > 1

# Element-wise logical statements
x <- c(-1, 0, 1)
```

```
[1] TRUE

[1] FALSE

[1] TRUE

[1] TRUE

[1] FALSE

[1] FALSE

[1] TRUE

[1] FALSE
```

Logical comparisons

```
x <- 1
x > 0
x > 2
x == 1
x != 2
!(x == x)
"s" == "S"
1 > 0 | 0 > 1
1 > 0 & 0 > 1

# Element-wise logical statements
x <- c(-1, 0, 1)
x < 0
```

```
[1] TRUE

[1] FALSE

[1] TRUE

[1] TRUE

[1] FALSE

[1] FALSE

[1] TRUE

[1] FALSE

[1]  TRUE FALSE FALSE
```

Logical comparisons

```
x <- 1
x > 0
x > 2
x == 1
x != 2
!(x == x)
"s" == "S"
1 > 0 | 0 > 1
1 > 0 & 0 > 1

# Element-wise logical statements
x <- c(-1, 0, 1)
x < 0
x == 0
```

```
[1] TRUE

[1] FALSE

[1] TRUE

[1] TRUE

[1] FALSE

[1] FALSE

[1] TRUE

[1] FALSE

[1]  TRUE FALSE FALSE

[1] FALSE  TRUE FALSE
```

Logical comparisons

```r
x <- 1
x > 0
x > 2
x == 1
x != 2
!(x == x)
"s" == "S"
1 > 0 | 0 > 1
1 > 0 & 0 > 1

# Element-wise logical statements
x <- c(-1, 0, 1)
x < 0
x == 0

# Quickly test is a value is contained in a set
1 %in% x
```

```
[1] TRUE

[1] FALSE

[1] TRUE

[1] TRUE

[1] FALSE

[1] FALSE

[1] TRUE

[1] FALSE

[1]  TRUE FALSE FALSE

[1] FALSE  TRUE FALSE

[1] TRUE
```

How do we use logical statements?

- One common way is through the `ifelse()` command?
- `ifelse(LOGICAL STATEMENT, DO IF TRUE, DO IF FALSE)`

```
x <- c(0, 1, 2, 3, NA)
ifelse(NA %in% x, "x contains a missing value", "x does not contain a missing value")
```

```
[1] "x contains a missing value"
```

```
x <- c(0, 1, 2, 3)
ifelse(NA %in% x, "x contains a missing value", "x does not contain a missing value")
```

```
[1] "x does not contain a missing value"
```

Data transformation with the dplyr package

- A toolbox for common data processing/manipulation operations to apply to tibbles
  - `glimpse()` to see the structure of a tibble
  - `arrange()` to sort data by columns
  - `filter()` and `slice()` to subset data by rows
  - `select()` to subset data by columns
  - `rename()` to rename columns
  - `mutate()` to add or change columns (or their values)
  - `summarize()` and `count()` to calculate summary statistics over rows of data
  - `group_by()` to perform operations within subsets of data
  - and many more...

Each dplyr function uses a similar structure

- `function(data, something_to_do_with_columns)`
- In base R, you often need to specify the dataset over and over to access the columns: `ds[ds$col1 == 1, 0]`
- In dplyr functions, the data argument lets you work with column names without ever using `$`. The data argument also lets you access the columns directly without using quoted expressions: `filter(data, col1 == 1)`
- Most dplyr functions return the entire tibble back as an output

Arrange (sorting data)

```
library(tidyverse)  #loads dplyr
ds <- starwars  #loads built-in star wars database
glimpse(ds)
```

```
Rows: 87
Columns: 14
$ name       <chr> "Luke Skywalker", "C-3PO", "R2-D2", "Darth Vader", "Leia Or…
$ height     <int> 172, 167, 96, 202, 150, 178, 165, 97, 183, 182, 188, 180, 2…
$ mass       <dbl> 77.0, 75.0, 32.0, 136.0, 49.0, 120.0, 75.0, 32.0, 84.0, 77.…
$ hair_color <chr> "blond", NA, NA, "none", "brown", "brown, grey", "brown", N…
$ skin_color <chr> "fair", "gold", "white, blue", "white", "light", "light", "…
$ eye_color  <chr> "blue", "yellow", "red", "yellow", "brown", "blue", "blue",…
$ birth_year <dbl> 19.0, 112.0, 33.0, 41.9, 19.0, 52.0, 47.0, NA, 24.0, 57.0, …
$ sex        <chr> "male", "none", "none", "male", "female", "male", "female",…
$ gender     <chr> "masculine", "masculine", "masculine", "masculine", "femini…
$ homeworld  <chr> "Tatooine", "Tatooine", "Naboo", "Tatooine", "Alderaan", "T…
$ species    <chr> "Human", "Droid", "Droid", "Human", "Human", "Human", "Huma…
$ films      <list> <"The Empire Strikes Back", "Revenge of the Sith", "Return…
$ vehicles   <list> <"Snowspeeder", "Imperial Speeder Bike">, <>, <>, <>, "Imp…
$ starships  <list> <"X-wing", "Imperial shuttle">, <>, <>, "TIE Advanced x1",…
```

Arrange (sorting data)

```
library(tidyverse)  #loads dplyr
ds <- starwars  #loads built-in star wars database
arrange(ds, name)
```

```
# A tibble: 87 × 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 mott… orange           41   male  mascu…
 2 Adi Ga…    184    50 none       dark        blue             NA   fema… femin…
 3 Anakin…    188    84 blond      fair        blue             41.9 male  mascu…
 4 Arvel …     NA    NA brown      fair        brown            NA   male  mascu…
 5 Ayla S…    178    55 none       blue        hazel            48   fema… femin…
 6 Bail P…    191    NA black      tan         brown            67   male  mascu…
 7 Barris…    166    50 black      yellow      blue             40   fema… femin…
 8 BB8         NA    NA none       none        black            NA   none  mascu…
 9 Ben Qu…    163    65 none       grey, gree… orange           NA   male  mascu…
10 Beru W…    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>
```

Arrange (sorting data)

```
library(tidyverse)  #loads dplyr
ds <- starwars  #loads built-in star wars database
arrange(ds, height)
```

```
# A tibble: 87 × 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 (sorting data)

```
library(tidyverse)  #loads dplyr
ds <- starwars  #loads built-in star wars database
arrange(ds, desc(height), mass)
```

```
# A tibble: 87 × 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 Chewba…    228   112 brown      unknown     blue            200   male  mascu…
 5 Roos T…    224    82 none       grey        orange           NA   male  mascu…
 6 Grievo…    216   159 none       brown, whi… green, y…        NA   male  mascu…
 7 Taun We    213    NA none       grey        black            NA   fema… femin…
 8 Tion M…    206    80 none       grey        black            NA   male  mascu…
 9 Rugor …    206    NA none       green       orange           NA   male  mascu…
10 Darth …    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>
```

Arrange (sorting data)

```
library(tidyverse)  #loads dplyr
ds <- starwars  #loads built-in star wars database
arrange(ds, eye_color, hair_color)
```

```
# A tibble: 87 × 14
   name   height  mass hair_color skin_color   eye_color birth_year sex   gender
   <chr>   <int> <dbl> <chr>      <chr>        <chr>           <dbl> <chr> <chr>
 1 Nien …    160    68 none       grey         black              NA male  mascu…
 2 Gasga…    122    NA none       white, blue  black              NA male  mascu…
 3 Kit F…    196    87 none       green        black              NA male  mascu…
 4 Plo K…    188    80 none       orange       black              22 male  mascu…
 5 Lama …    229    88 none       grey         black              NA male  mascu…
 6 Taun …    213    NA none       grey         black              NA fema… femin…
 7 Shaak…    178    57 none       red, blue, … black              NA fema… femin…
 8 Tion …    206    80 none       grey         black              NA male  mascu…
 9 BB8        NA    NA none       none         black              NA none  mascu…
10 Greedo    173    74 <NA>       green        black              44 male  mascu…
# … with 77 more rows, and 5 more variables: homeworld <chr>, species <chr>,
#   films <list>, vehicles <list>, starships <list>
```

Filter (subsetting rows)

```
filter(ds, height < 100)
```

```
# A tibble: 7 × 14
  name      height  mass hair_color skin_color eye_color birth_year sex   gender
  <chr>      <int> <dbl> <chr>      <chr>      <chr>           <dbl> <chr> <chr>
1 R2-D2         96    32 <NA>       white, bl… red                33 none  mascu…
2 R5-D4         97    32 <NA>       white, red red                NA none  mascu…
3 Yoda          66    17 white      green      brown             896 male  mascu…
4 Wicket S…     88    20 brown      brown      brown               8 male  mascu…
5 Dud Bolt      94    45 none       blue, grey yellow             NA male  mascu…
6 Ratts Ty…     79    15 none       grey, blue unknown            NA male  mascu…
7 R4-P17        96    NA none       silver, r… red, blue          NA none  femin…
# … with 5 more variables: homeworld <chr>, species <chr>, films <list>,
#   vehicles <list>, starships <list>
```

Filter (subsetting rows)

```
filter(ds, name == "Yoda")
```

```
# A tibble: 1 × 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  masculine
# … with 5 more variables: homeworld <chr>, species <chr>, films <list>,
#   vehicles <list>, starships <list>
```

Filter (subsetting rows)

```
filter(ds, is.na(hair_color))
```

```
# A tibble: 5 × 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, blue 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 …    175  1358 <NA>       green-tan,… orange            600 herma… mascu…
# … with 5 more variables: homeworld <chr>, species <chr>, films <list>,
#   vehicles <list>, starships <list>
```

Filter (subsetting rows)

```
filter(ds, height > 100, height < 150)
```

```
# A tibble: 3 × 14
  name    height  mass hair_color skin_color  eye_color birth_year sex   gender
  <chr>    <int> <dbl> <chr>      <chr>       <chr>           <dbl> <chr> <chr>
1 Watto      137    NA black      blue, grey  yellow             NA male  mascul…
2 Sebulba    112    40 none       grey, red   orange             NA male  mascul…
3 Gasgano    122    NA none       white, blue black              NA male  mascul…
# … with 5 more variables: homeworld <chr>, species <chr>, films <list>,
#   vehicles <list>, starships <list>
```

Filter (subsetting rows)

```
filter(ds, eye_color %in% c("blue","brown"))
```

```
# A tibble: 40 × 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 Leia Or…    150    49 brown      light      brown            19   fema… femin…
 3 Owen La…    178   120 brown, gr… light      blue             52   male  mascu…
 4 Beru Wh…    165    75 brown      light      blue             47   fema… femin…
 5 Biggs D…    183    84 black      light      brown            24   male  mascu…
 6 Anakin …    188    84 blond      fair       blue             41.9 male  mascu…
 7 Wilhuff…    180    NA auburn, g… fair       blue             64   male  mascu…
 8 Chewbac…    228   112 brown      unknown    blue            200   male  mascu…
 9 Han Solo    180    80 brown      fair       brown            29   male  mascu…
10 Jek Ton…    180   110 brown      fair       blue             NA   male  mascu…
# … with 30 more rows, and 5 more variables: homeworld <chr>, species <chr>,
#   films <list>, vehicles <list>, starships <list>
```

Filter (subsetting rows)

```
filter(ds, !(eye_color %in% c("blue","brown")))
```

```
# A tibble: 47 × 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 Darth …    202   136 none        white      yellow           41.9 male  mascu…
 4 R5-D4       97    32 <NA>        white, red red              NA   none  mascu…
 5 Obi-Wa…    182    77 auburn, wh… fair       blue-gray        57   male  mascu…
 6 Greedo     173    74 <NA>        green      black            44   male  mascu…
 7 Jabba …    175  1358 <NA>        green-tan… orange          600   herm… mascu…
 8 Wedge …    170    77 brown       fair       hazel            21   male  mascu…
 9 Palpat…    170    75 grey        pale       yellow           82   male  mascu…
10 IG-88      200   140 none        metal      red              15   none  mascu…
# … with 37 more rows, and 5 more variables: homeworld <chr>, species <chr>,
#   films <list>, vehicles <list>, starships <list>
```

Slice (subsetting rows by position)

```
slice(ds, 1:5)
```

```
# A tibble: 5 × 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, blue 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…
# … with 5 more variables: homeworld <chr>, species <chr>, films <list>,
#   vehicles <list>, starships <list>
```

Slice (subsetting rows by position)

```
slice_head(ds, n = 5)
```

```
# A tibble: 5 × 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, blue 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…
# … with 5 more variables: homeworld <chr>, species <chr>, films <list>,
#   vehicles <list>, starships <list>
```

Slice (subsetting rows by position)

```
slice_tail(ds, n = 4)
```

```
# A tibble: 4 × 14
  name     height  mass hair_color skin_color eye_color birth_year sex    gender
  <chr>     <int> <dbl> <chr>      <chr>      <chr>           <dbl> <chr>  <chr>
1 Poe Dam…     NA    NA brown      light      brown              NA male   mascu…
2 BB8          NA    NA none       none       black              NA none   mascu…
3 Captain…     NA    NA unknown    unknown    unknown            NA <NA>   <NA>
4 Padmé A…    165    45 brown      light      brown              46 female femin…
# … with 5 more variables: homeworld <chr>, species <chr>, films <list>,
#   vehicles <list>, starships <list>
```

Slice (subsetting rows by position)

```
slice_sample(ds, n = 2)
```

```
# A tibble: 2 × 14
  name     height  mass hair_color skin_color eye_color birth_year sex   gender
  <chr>     <int> <dbl> <chr>      <chr>      <chr>           <dbl> <chr> <chr>
1 Mace Wi…    188    84 none       dark       brown              72 male  mascul…
2 Sly Moo…    178    48 none       pale       white              NA <NA>  <NA>
# … with 5 more variables: homeworld <chr>, species <chr>, films <list>,
#   vehicles <list>, starships <list>
```

Slice (subsetting rows by position)

```
slice_sample(ds, n = 2)
```

```
# A tibble: 2 × 14
  name     height  mass hair_color skin_color eye_color birth_year sex   gender
  <chr>     <int> <dbl> <chr>      <chr>      <chr>           <dbl> <chr> <chr>
1 Ki-Adi-…    198    82 white      pale       yellow             92 male  mascul…
2 Sebulba     112    40 none       grey, red  orange             NA male  mascul…
# … with 5 more variables: homeworld <chr>, species <chr>, films <list>,
#   vehicles <list>, starships <list>
```

Slice (subsetting rows by position)

```
slice_min(ds, height, n = 3)
```

```
# A tibble: 3 × 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 Ty…     79    15 none       grey, blue unknown            NA male  mascu…
3 Wicket S…     88    20 brown      brown      brown               8 male  mascu…
# … with 5 more variables: homeworld <chr>, species <chr>, films <list>,
#   vehicles <list>, starships <list>
```

Select (subsetting columns)

```
select(ds, name)
```

```
# A tibble: 87 × 1
   name
   <chr>
 1 Luke Skywalker
 2 C-3PO
 3 R2-D2
 4 Darth Vader
 5 Leia Organa
 6 Owen Lars
 7 Beru Whitesun lars
 8 R5-D4
 9 Biggs Darklighter
10 Obi-Wan Kenobi
# … with 77 more rows
```

Select (subsetting columns)

```
select(ds, name, height, mass)
```

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

Select (subsetting columns)

```
select(ds, c("name", "height", "mass"))
```

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

Select (subsetting columns)

```
select(ds, name:eye_color)
```

```
# A tibble: 87 × 6
   name               height  mass hair_color    skin_color  eye_color
   <chr>               <int> <dbl> <chr>         <chr>       <chr>
 1 Luke Skywalker        172    77 blond         fair        blue
 2 C-3PO                 167    75 <NA>          gold        yellow
 3 R2-D2                  96    32 <NA>          white, blue red
 4 Darth Vader           202   136 none          white       yellow
 5 Leia Organa           150    49 brown         light       brown
 6 Owen Lars             178   120 brown, grey   light       blue
 7 Beru Whitesun lars    165    75 brown         light       blue
 8 R5-D4                  97    32 <NA>          white, red  red
 9 Biggs Darklighter     183    84 black         light       brown
10 Obi-Wan Kenobi        182    77 auburn, white fair        blue-gray
# … with 77 more rows
```

Select (subsetting columns)

```
select(ds, -(eye_color:starships))
```

```
# A tibble: 87 × 5
   name               height  mass hair_color    skin_color
   <chr>               <int> <dbl> <chr>         <chr>
 1 Luke Skywalker        172    77 blond         fair
 2 C-3PO                 167    75 <NA>          gold
 3 R2-D2                  96    32 <NA>          white, blue
 4 Darth Vader           202   136 none          white
 5 Leia Organa           150    49 brown         light
 6 Owen Lars             178   120 brown, grey   light
 7 Beru Whitesun lars    165    75 brown         light
 8 R5-D4                  97    32 <NA>          white, red
 9 Biggs Darklighter     183    84 black         light
10 Obi-Wan Kenobi        182    77 auburn, white fair
# … with 77 more rows
```

Select (subsetting columns)

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

```
# A tibble: 87 × 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 (subsetting columns)

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

```
# A tibble: 87 × 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 (subsetting columns)

```
select(ds, where(is.numeric))
```

```
# A tibble: 87 × 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
 7    165    75       47
 8     97    32       NA
 9    183    84       24
10    182    77       57
# … with 77 more rows
```

Select (subsetting columns)

```
select(ds, where(is.character))
```

```
# A tibble: 87 × 8
   name         hair_color   skin_color eye_color sex   gender  homeworld species
   <chr>        <chr>        <chr>      <chr>     <chr> <chr>   <chr>     <chr>
 1 Luke Skywa… blond        fair       blue      male  mascul… Tatooine  Human
 2 C-3PO       <NA>         gold       yellow    none  mascul… Tatooine  Droid
 3 R2-D2       <NA>         white, bl… red       none  mascul… Naboo     Droid
 4 Darth Vader none         white      yellow    male  mascul… Tatooine  Human
 5 Leia Organa brown        light      brown     fema… femini… Alderaan  Human
 6 Owen Lars   brown, grey  light      blue      male  mascul… Tatooine  Human
 7 Beru White… brown        light      blue      fema… femini… Tatooine  Human
 8 R5-D4       <NA>         white, red red       none  mascul… Tatooine  Droid
 9 Biggs Dark… black        light      brown     male  mascul… Tatooine  Human
10 Obi-Wan Ke… auburn, whi… fair       blue-gray male  mascul… Stewjon   Human
# … with 77 more rows
```

What's going on here?

```
select(ds, name, height, eye_color)
```

```
# A tibble: 87 × 3
   name               height eye_color
   <chr>               <int> <chr>
 1 Luke Skywalker        172 blue
 2 C-3PO                 167 yellow
 3 R2-D2                  96 red
 4 Darth Vader           202 yellow
 5 Leia Organa           150 brown
 6 Owen Lars             178 blue
 7 Beru Whitesun lars    165 blue
 8 R5-D4                  97 red
 9 Biggs Darklighter     183 brown
10 Obi-Wan Kenobi        182 blue-gray
# … with 77 more rows
```

What's going on here?

```
select(ds, name, height, eye_color)
filter(ds, height < 70)
```

```
# A tibble: 87 × 3
   name               height eye_color
   <chr>               <int> <chr>
 1 Luke Skywalker        172 blue
 2 C-3PO                 167 yellow
 3 R2-D2                  96 red
 4 Darth Vader           202 yellow
 5 Leia Organa           150 brown
 6 Owen Lars             178 blue
 7 Beru Whitesun lars    165 blue
 8 R5-D4                  97 red
 9 Biggs Darklighter     183 brown
10 Obi-Wan Kenobi        182 blue-gray
# … with 77 more rows


# A tibble: 1 × 14
  name  height  mass hair_color skin_color eye_color birth_
  <chr>  <int> <dbl> <chr>      <chr>      <chr>           <
1 Yoda      66    17 white      green      brown
# … with 5 more variables: homeworld <chr>, species <chr>,
#   vehicles <list>, starships <list>
```

What's going on here?

```
select(ds, name, height, eye_color)
filter(ds, height < 70)
ds
```

```
# A tibble: 87 × 3
   name               height eye_color
   <chr>               <int> <chr>
 1 Luke Skywalker        172 blue
 2 C-3PO                 167 yellow
 3 R2-D2                  96 red
 4 Darth Vader           202 yellow
 5 Leia Organa           150 brown
 6 Owen Lars             178 blue
 7 Beru Whitesun lars    165 blue
 8 R5-D4                  97 red
 9 Biggs Darklighter     183 brown
10 Obi-Wan Kenobi        182 blue-gray
# … with 77 more rows

# A tibble: 1 × 14
  name  height  mass hair_color skin_color eye_color birth_
  <chr>  <int> <dbl> <chr>      <chr>      <chr>           <
1 Yoda      66    17 white      green      brown
# … with 5 more variables: homeworld <chr>, species <chr>,
#   vehicles <list>, starships <list>

# A tibble: 87 × 14
   name    height  mass hair_color  skin_color eye_color bi
   <chr>    <int> <dbl> <chr>       <chr>      <chr>
 1 Luke S…    172    77 blond       fair       blue
 2 C-3PO      167    75 <NA>        gold       yellow
 3 R2-D2       96    32 <NA>        white, bl… red
 4 Darth …    202   136 none        white      yellow
```

Reassign the transformations back to the tibble

```
ds <- select(ds, name, height, eye_color)

ds
```

```
# A tibble: 87 × 3
   name               height eye_color
   <chr>               <int> <chr>
 1 Luke Skywalker        172 blue
 2 C-3PO                 167 yellow
 3 R2-D2                  96 red
 4 Darth Vader           202 yellow
 5 Leia Organa           150 brown
 6 Owen Lars             178 blue
 7 Beru Whitesun lars    165 blue
 8 R5-D4                  97 red
 9 Biggs Darklighter     183 brown
10 Obi-Wan Kenobi        182 blue-gray
# … with 77 more rows
```

Reassign the transformations back to the tibble

```
ds <- select(ds, name, height, eye_color)
ds <- arrange(ds, height, eye_color)

ds
```

```
# A tibble: 87 × 3
   name                  height eye_color
   <chr>                  <int> <chr>
 1 Yoda                      66 brown
 2 Ratts Tyerell             79 unknown
 3 Wicket Systri Warrick     88 brown
 4 Dud Bolt                  94 yellow
 5 R2-D2                     96 red
 6 R4-P17                    96 red, blue
 7 R5-D4                     97 red
 8 Sebulba                  112 orange
 9 Gasgano                  122 black
10 Watto                    137 yellow
# … with 77 more rows
```

Reassign the transformations back to the tibble

```
ds <- select(ds, name, height, eye_color)
ds <- arrange(ds, height, eye_color)
ds <- filter(ds, height < 70)

ds
```

```
# A tibble: 1 × 3
  name  height eye_color
  <chr>  <int> <chr>
1 Yoda       66 brown
```

Not a good strategy

```
ds <- starwars
```

Not a good strategy

```
ds <- starwars
ds_name_height_eye_color <- select(ds, name, height, eye_color)
```

Not a good strategy

```
ds <- starwars
ds_name_height_eye_color <- select(ds, name, height, eye_color)
ds_sorted <- arrange(ds_name_height_eye_color, height, eye_color)
```

Not a good strategy

```
ds <- starwars
ds_name_height_eye_color <- select(ds, name, height, eye_color)
ds_sorted <- arrange(ds_name_height_eye_color, height, eye_color)
ds_sorted_filtered <- filter(ds_sorted, height < 70)
```

Not a good strategy

```
ds <- starwars
ds_name_height_eye_color <- select(ds, name, height, eye_color)
ds_sorted <- arrange(ds_name_height_eye_color, height, eye_color)
ds_sorted_filtered <- filter(ds_sorted, height < 70)
ds_sorted_filtered
```

```
# A tibble: 1 × 3
  name  height eye_color
  <chr>  <int> <chr>
1 Yoda      66 brown
```

Introducing the pipe operator: %>%

```
ds <- starwars

ds
```

```
# A tibble: 87 × 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 S…    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 …    202   136 none        white      yellow           41.9 male  mascu…
 5 Leia O…    150    49 brown       light      brown            19   fema… femin…
 6 Owen L…    178   120 brown, grey light      blue             52   male  mascu…
 7 Beru W…    165    75 brown       light      blue             47   fema… femin…
 8 R5-D4       97    32 <NA>        white, red red              NA   none  mascu…
 9 Biggs …    183    84 black       light      brown            24   male  mascu…
10 Obi-Wa…    182    77 auburn, wh… fair       blue-gray        57   male  mascu…
# … with 77 more rows, and 5 more variables: homeworld <chr>, species <chr>,
#   films <list>, vehicles <list>, starships <list>
```

Introducing the pipe operator: %>%

```
ds <- starwars
ds <- ds %>% select(name, height, eye_color)

ds
```

```
# A tibble: 87 × 3
   name               height eye_color
   <chr>               <int> <chr>
 1 Luke Skywalker        172 blue
 2 C-3PO                 167 yellow
 3 R2-D2                  96 red
 4 Darth Vader           202 yellow
 5 Leia Organa           150 brown
 6 Owen Lars             178 blue
 7 Beru Whitesun lars    165 blue
 8 R5-D4                  97 red
 9 Biggs Darklighter     183 brown
10 Obi-Wan Kenobi        182 blue-gray
# … with 77 more rows
```

Introducing the pipe operator: %>%

```
ds <- starwars
ds <- ds %>% select(name, height, eye_color)

ds <- ds

ds
```

```
# A tibble: 87 × 3
   name               height eye_color
   <chr>               <int> <chr>
 1 Luke Skywalker        172 blue
 2 C-3PO                 167 yellow
 3 R2-D2                  96 red
 4 Darth Vader           202 yellow
 5 Leia Organa           150 brown
 6 Owen Lars             178 blue
 7 Beru Whitesun lars    165 blue
 8 R5-D4                  97 red
 9 Biggs Darklighter     183 brown
10 Obi-Wan Kenobi        182 blue-gray
# … with 77 more rows
```

Introducing the pipe operator: %>%

```
ds <- starwars
ds <- ds %>% select(name, height, eye_color)

ds <- ds %>%
  select(name, eye_color)

ds
```

```
# A tibble: 87 × 2
   name               eye_color
   <chr>              <chr>
 1 Luke Skywalker     blue
 2 C-3PO              yellow
 3 R2-D2              red
 4 Darth Vader        yellow
 5 Leia Organa        brown
 6 Owen Lars          blue
 7 Beru Whitesun lars blue
 8 R5-D4              red
 9 Biggs Darklighter  brown
10 Obi-Wan Kenobi     blue-gray
# … with 77 more rows
```

Introducing the pipe operator: %>%

```
ds <- starwars
ds <- ds %>% select(name, height, eye_color)

ds <- ds %>%
  select(name, eye_color) %>%
  arrange(eye_color)

ds
```

```
# A tibble: 87 × 2
   name       eye_color
   <chr>      <chr>
 1 Greedo     black
 2 Nien Nunb  black
 3 Gasgano    black
 4 Kit Fisto  black
 5 Plo Koon   black
 6 Lama Su    black
 7 Taun We    black
 8 Shaak Ti   black
 9 Tion Medon black
10 BB8        black
# … with 77 more rows
```

Introducing the pipe operator: %>%

```
ds <- starwars
ds <- ds %>% select(name, height, eye_color)

ds <- ds %>%
  select(name, eye_color) %>%
  arrange(eye_color) %>%
  filter(eye_color == "blue")

ds
```

```
# A tibble: 19 × 2
   name                eye_color
   <chr>               <chr>
 1 Luke Skywalker      blue
 2 Owen Lars           blue
 3 Beru Whitesun lars  blue
 4 Anakin Skywalker    blue
 5 Wilhuff Tarkin      blue
 6 Chewbacca           blue
 7 Jek Tono Porkins    blue
 8 Lobot               blue
 9 Mon Mothma          blue
10 Qui-Gon Jinn        blue
11 Finis Valorum       blue
12 Ric Olié            blue
13 Adi Gallia          blue
14 Mas Amedda          blue
15 Cliegg Lars         blue
16 Luminara Unduli     blue
17 Barriss Offee       blue
```

This is not the most intuitive thing at first

All the options below are equivalent, but I stick with `ds <- ds %>%` because it's the most common to see

```
ds <- ds %>% select(height) %>% slice_tail(n = 5)

ds %>% select(height) %>% slice_tail(n = 5) -> ds

ds <- select(ds, height) %>% slice_tail(n = 5)

ds <- slice_tail(select(ds, height), n = 5)
```

Rename

- Consistent and clear names help prevent errors
    - Avoid names like "dv1", "dv2" that are difficult to remember
    - Auto-complete within RStudio means that you don't usually need to type out longer names
- Using rename can help clean up messy names from other sources
    - Avoid names that contain spaces or start with numbers
    - Force names to use a similar format like snake_case or camelCase
- Installing and loading the `janitor` package opens up some helpful renaming options

Rename columns

```
library(janitor)
# Switching to built-in iris data set since starwars has good names
iris
```

```
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
4            4.6         3.1          1.5         0.2     setosa
5            5.0         3.6          1.4         0.2     setosa
6            5.4         3.9          1.7         0.4     setosa
7            4.6         3.4          1.4         0.3     setosa
8            5.0         3.4          1.5         0.2     setosa
9            4.4         2.9          1.4         0.2     setosa
10           4.9         3.1          1.5         0.1     setosa
11           5.4         3.7          1.5         0.2     setosa
12           4.8         3.4          1.6         0.2     setosa
13           4.8         3.0          1.4         0.1     setosa
14           4.3         3.0          1.1         0.1     setosa
15           5.8         4.0          1.2         0.2     setosa
16           5.7         4.4          1.5         0.4     setosa
17           5.4         3.9          1.3         0.4     setosa
18           5.1         3.5          1.4         0.3     setosa
19           5.7         3.8          1.7         0.3     setosa
20           5.1         3.8          1.5         0.3     setosa
21           5.4         3.4          1.7         0.2     setosa
22           5.1         3.7          1.5         0.4     setosa
23           4.6         3.6          1.0         0.2     setosa
24           5.1         3.3          1.7         0.5     setosa
25           4.8         3.4          1.9         0.2     setosa
```

Rename columns

```
library(janitor)
# Switching to built-in iris data set since starwars has good names
iris %>% rename(sepal_length = Sepal.Length)
```

```
    sepal_length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
4            4.6         3.1          1.5         0.2     setosa
5            5.0         3.6          1.4         0.2     setosa
6            5.4         3.9          1.7         0.4     setosa
7            4.6         3.4          1.4         0.3     setosa
8            5.0         3.4          1.5         0.2     setosa
9            4.4         2.9          1.4         0.2     setosa
10           4.9         3.1          1.5         0.1     setosa
11           5.4         3.7          1.5         0.2     setosa
12           4.8         3.4          1.6         0.2     setosa
13           4.8         3.0          1.4         0.1     setosa
14           4.3         3.0          1.1         0.1     setosa
15           5.8         4.0          1.2         0.2     setosa
16           5.7         4.4          1.5         0.4     setosa
17           5.4         3.9          1.3         0.4     setosa
18           5.1         3.5          1.4         0.3     setosa
19           5.7         3.8          1.7         0.3     setosa
20           5.1         3.8          1.5         0.3     setosa
21           5.4         3.4          1.7         0.2     setosa
22           5.1         3.7          1.5         0.4     setosa
23           4.6         3.6          1.0         0.2     setosa
24           5.1         3.3          1.7         0.5     setosa
25           4.8         3.4          1.9         0.2     setosa
```

Rename columns

```
library(janitor)
# Switching to built-in iris data set since starwars has good names
iris %>% rename(sepal_length = Sepal.Length, sepal_width = Sepal.Width, petal_length =
```

```
    sepal_length sepal_width petal_length petal_width    species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
4            4.6         3.1          1.5         0.2     setosa
5            5.0         3.6          1.4         0.2     setosa
6            5.4         3.9          1.7         0.4     setosa
7            4.6         3.4          1.4         0.3     setosa
8            5.0         3.4          1.5         0.2     setosa
9            4.4         2.9          1.4         0.2     setosa
10           4.9         3.1          1.5         0.1     setosa
11           5.4         3.7          1.5         0.2     setosa
12           4.8         3.4          1.6         0.2     setosa
13           4.8         3.0          1.4         0.1     setosa
14           4.3         3.0          1.1         0.1     setosa
15           5.8         4.0          1.2         0.2     setosa
16           5.7         4.4          1.5         0.4     setosa
17           5.4         3.9          1.3         0.4     setosa
18           5.1         3.5          1.4         0.3     setosa
19           5.7         3.8          1.7         0.3     setosa
20           5.1         3.8          1.5         0.3     setosa
21           5.4         3.4          1.7         0.2     setosa
22           5.1         3.7          1.5         0.4     setosa
23           4.6         3.6          1.0         0.2     setosa
24           5.1         3.3          1.7         0.5     setosa
25           4.8         3.4          1.9         0.2     setosa
```

Rename columns

```
library(janitor)
# Switching to built-in iris data set since starwars has good names
iris %>% rename_with(toupper)
```

```
    SEPAL.LENGTH SEPAL.WIDTH PETAL.LENGTH PETAL.WIDTH    SPECIES
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
4            4.6         3.1          1.5         0.2     setosa
5            5.0         3.6          1.4         0.2     setosa
6            5.4         3.9          1.7         0.4     setosa
7            4.6         3.4          1.4         0.3     setosa
8            5.0         3.4          1.5         0.2     setosa
9            4.4         2.9          1.4         0.2     setosa
10           4.9         3.1          1.5         0.1     setosa
11           5.4         3.7          1.5         0.2     setosa
12           4.8         3.4          1.6         0.2     setosa
13           4.8         3.0          1.4         0.1     setosa
14           4.3         3.0          1.1         0.1     setosa
15           5.8         4.0          1.2         0.2     setosa
16           5.7         4.4          1.5         0.4     setosa
17           5.4         3.9          1.3         0.4     setosa
18           5.1         3.5          1.4         0.3     setosa
19           5.7         3.8          1.7         0.3     setosa
20           5.1         3.8          1.5         0.3     setosa
21           5.4         3.4          1.7         0.2     setosa
22           5.1         3.7          1.5         0.4     setosa
23           4.6         3.6          1.0         0.2     setosa
24           5.1         3.3          1.7         0.5     setosa
25           4.8         3.4          1.9         0.2     setosa
```

Rename columns

```
library(janitor)
# Switching to built-in iris data set since starwars has good names
iris %>% rename_with(tolower, starts_with("Petal"))
```

```
    Sepal.Length Sepal.Width petal.length petal.width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
4            4.6         3.1          1.5         0.2     setosa
5            5.0         3.6          1.4         0.2     setosa
6            5.4         3.9          1.7         0.4     setosa
7            4.6         3.4          1.4         0.3     setosa
8            5.0         3.4          1.5         0.2     setosa
9            4.4         2.9          1.4         0.2     setosa
10           4.9         3.1          1.5         0.1     setosa
11           5.4         3.7          1.5         0.2     setosa
12           4.8         3.4          1.6         0.2     setosa
13           4.8         3.0          1.4         0.1     setosa
14           4.3         3.0          1.1         0.1     setosa
15           5.8         4.0          1.2         0.2     setosa
16           5.7         4.4          1.5         0.4     setosa
17           5.4         3.9          1.3         0.4     setosa
18           5.1         3.5          1.4         0.3     setosa
19           5.7         3.8          1.7         0.3     setosa
20           5.1         3.8          1.5         0.3     setosa
21           5.4         3.4          1.7         0.2     setosa
22           5.1         3.7          1.5         0.4     setosa
23           4.6         3.6          1.0         0.2     setosa
24           5.1         3.3          1.7         0.5     setosa
25           4.8         3.4          1.9         0.2     setosa
```

Rename columns

```
library(janitor)
# Switching to built-in iris data set since starwars has good names
iris %>% clean_names()
```

```
    sepal_length sepal_width petal_length petal_width    species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
4            4.6         3.1          1.5         0.2     setosa
5            5.0         3.6          1.4         0.2     setosa
6            5.4         3.9          1.7         0.4     setosa
7            4.6         3.4          1.4         0.3     setosa
8            5.0         3.4          1.5         0.2     setosa
9            4.4         2.9          1.4         0.2     setosa
10           4.9         3.1          1.5         0.1     setosa
11           5.4         3.7          1.5         0.2     setosa
12           4.8         3.4          1.6         0.2     setosa
13           4.8         3.0          1.4         0.1     setosa
14           4.3         3.0          1.1         0.1     setosa
15           5.8         4.0          1.2         0.2     setosa
16           5.7         4.4          1.5         0.4     setosa
17           5.4         3.9          1.3         0.4     setosa
18           5.1         3.5          1.4         0.3     setosa
19           5.7         3.8          1.7         0.3     setosa
20           5.1         3.8          1.5         0.3     setosa
21           5.4         3.4          1.7         0.2     setosa
22           5.1         3.7          1.5         0.4     setosa
23           4.6         3.6          1.0         0.2     setosa
24           5.1         3.3          1.7         0.5     setosa
25           4.8         3.4          1.9         0.2     setosa
```

Rename columns

```
library(janitor)
# Switching to built-in iris data set since starwars has good names
iris %>% clean_names("small_camel")
```

```
    sepalLength sepalWidth petalLength petalWidth    species
1           5.1         3.5          1.4         0.2     setosa
2           4.9         3.0          1.4         0.2     setosa
3           4.7         3.2          1.3         0.2     setosa
4           4.6         3.1          1.5         0.2     setosa
5           5.0         3.6          1.4         0.2     setosa
6           5.4         3.9          1.7         0.4     setosa
7           4.6         3.4          1.4         0.3     setosa
8           5.0         3.4          1.5         0.2     setosa
9           4.4         2.9          1.4         0.2     setosa
10          4.9         3.1          1.5         0.1     setosa
11          5.4         3.7          1.5         0.2     setosa
12          4.8         3.4          1.6         0.2     setosa
13          4.8         3.0          1.4         0.1     setosa
14          4.3         3.0          1.1         0.1     setosa
15          5.8         4.0          1.2         0.2     setosa
16          5.7         4.4          1.5         0.4     setosa
17          5.4         3.9          1.3         0.4     setosa
18          5.1         3.5          1.4         0.3     setosa
19          5.7         3.8          1.7         0.3     setosa
20          5.1         3.8          1.5         0.3     setosa
21          5.4         3.4          1.7         0.2     setosa
22          5.1         3.7          1.5         0.4     setosa
23          4.6         3.6          1.0         0.2     setosa
24          5.1         3.3          1.7         0.5     setosa
25          4.8         3.4          1.9         0.2     setosa
```

Mutate and summarize

- Mutate adds/modifies columns; dataset remains the same size
- Summarize collapses the data set down (by default to 1 row) to calculate summary statistics
- Both functions use a similar form:
  - `mutate(data, new_variable = expression)`
  - `summarize(data, mean = mean(var))`
- Like all other dplyr functions, they return tibbles, so for mutate to "stick" it needs to be assigned back to itself:
  - `data <- mutate(data, var = round(var))`
  - `data <- mutate(data, var_r = round(var))`
- Since summarize collapses to a new level, you might not want to assign it to overwrite the original data set!

Mutate (add or modify columns)

```
ds <- starwars %>% select(name, mass, height, hair_color)

ds
```

```
# A tibble: 87 × 4
   name                mass height hair_color
   <chr>              <dbl>  <int> <chr>
 1 Luke Skywalker        77    172 blond
 2 C-3PO                 75    167 <NA>
 3 R2-D2                 32     96 <NA>
 4 Darth Vader          136    202 none
 5 Leia Organa           49    150 brown
 6 Owen Lars            120    178 brown, grey
 7 Beru Whitesun lars    75    165 brown
 8 R5-D4                 32     97 <NA>
 9 Biggs Darklighter     84    183 black
10 Obi-Wan Kenobi        77    182 auburn, white
# … with 77 more rows
```

Mutate (add or modify columns)

```
ds <- starwars %>% select(name, mass, height, hair_color)
ds <- ds %>% mutate(in_movie = "yes")

ds
```

```
# A tibble: 87 × 5
   name                 mass height hair_color    in_movie
   <chr>               <dbl>  <int> <chr>         <chr>
 1 Luke Skywalker         77    172 blond         yes
 2 C-3PO                  75    167 <NA>          yes
 3 R2-D2                  32     96 <NA>          yes
 4 Darth Vader           136    202 none          yes
 5 Leia Organa            49    150 brown         yes
 6 Owen Lars             120    178 brown, grey   yes
 7 Beru Whitesun lars     75    165 brown         yes
 8 R5-D4                  32     97 <NA>          yes
 9 Biggs Darklighter      84    183 black         yes
10 Obi-Wan Kenobi         77    182 auburn, white yes
# … with 77 more rows
```

Mutate (add or modify columns)

```
ds <- starwars %>% select(name, mass, height, hair_color)
ds <- ds %>% mutate(in_movie = "yes")
ds <- ds %>% mutate(height_m = height/100,
                    bmi = mass/(height_m^2),
                    bmi = round(bmi))

ds
```

```
# A tibble: 87 × 7
   name               mass height hair_color    in_movie height_m   bmi
   <chr>             <dbl>  <int> <chr>         <chr>       <dbl> <dbl>
 1 Luke Skywalker       77    172 blond         yes          1.72    26
 2 C-3PO                75    167 <NA>          yes          1.67    27
 3 R2-D2                32     96 <NA>          yes          0.96    35
 4 Darth Vader         136    202 none          yes          2.02    33
 5 Leia Organa          49    150 brown         yes          1.5     22
 6 Owen Lars           120    178 brown, grey   yes          1.78    38
 7 Beru Whitesun lars   75    165 brown         yes          1.65    28
 8 R5-D4                32     97 <NA>          yes          0.97    34
 9 Biggs Darklighter    84    183 black         yes          1.83    25
10 Obi-Wan Kenobi       77    182 auburn, white yes          1.82    23
# … with 77 more rows
```

Mutate (add or modify columns)

```
ds <- starwars %>% select(name, mass, height, hair_color)
ds <- ds %>% mutate(in_movie = "yes")
ds <- ds %>% mutate(height_m = height/100,
                    bmi = mass/(height_m^2),
                    bmi = round(bmi)) %>%
             arrange(desc(bmi))

ds
```

```
# A tibble: 87 × 7
   name                  mass height hair_color  in_movie height_m   bmi
   <chr>                <dbl>  <int> <chr>       <chr>       <dbl> <dbl>
 1 Jabba Desilijic Tiure  1358    175 <NA>        yes          1.75   443
 2 Dud Bolt                 45     94 none        yes          0.94    51
 3 Yoda                     17     66 white       yes          0.66    39
 4 Owen Lars               120    178 brown, grey yes          1.78    38
 5 R2-D2                    32     96 <NA>        yes          0.96    35
 6 IG-88                   140    200 none        yes          2       35
 7 R5-D4                    32     97 <NA>        yes          0.97    34
 8 Jek Tono Porkins        110    180 brown       yes          1.8     34
 9 Grievous                159    216 none        yes          2.16    34
10 Darth Vader             136    202 none        yes          2.02    33
# … with 77 more rows
```

Mutate (add or modify columns)

```
ds <- starwars %>% select(name, mass, height, hair_color)
ds <- ds %>% mutate(in_movie = "yes")
ds <- ds %>% mutate(height_m = height/100,
                    bmi = mass/(height_m^2),
                    bmi = round(bmi)) %>%
             arrange(desc(bmi))

ds <- ds

ds
```

```
# A tibble: 87 × 7
   name                  mass height hair_color  in_movie height_m   bmi
   <chr>                <dbl>  <int> <chr>       <chr>       <dbl> <dbl>
 1 Jabba Desilijic Tiure  1358    175 <NA>        yes          1.75   443
 2 Dud Bolt                 45     94 none        yes          0.94    51
 3 Yoda                     17     66 white       yes          0.66    39
 4 Owen Lars               120    178 brown, grey yes          1.78    38
 5 R2-D2                    32     96 <NA>        yes          0.96    35
 6 IG-88                   140    200 none        yes          2       35
 7 R5-D4                    32     97 <NA>        yes          0.97    34
 8 Jek Tono Porkins        110    180 brown       yes          1.8     34
 9 Grievous                159    216 none        yes          2.16    34
10 Darth Vader             136    202 none        yes          2.02    33
# … with 77 more rows
```

Mutate (add or modify columns)

```
ds <- starwars %>% select(name, mass, height, hair_color)
ds <- ds %>% mutate(in_movie = "yes")
ds <- ds %>% mutate(height_m = height/100,
                    bmi = mass/(height_m^2),
                    bmi = round(bmi)) %>%
            arrange(desc(bmi))

ds <- ds %>%
  filter(hair_color %in% c("blond", NA))

ds
```

```
# A tibble: 8 × 7
  name                 mass height hair_color in_movie height_m   bmi
  <chr>               <dbl>  <int> <chr>      <chr>        <dbl> <dbl>
1 Jabba Desilijic Tiure  1358    175 <NA>       yes           1.75   443
2 R2-D2                    32     96 <NA>       yes           0.96    35
3 R5-D4                    32     97 <NA>       yes           0.97    34
4 C-3PO                    75    167 <NA>       yes           1.67    27
5 Luke Skywalker           77    172 blond      yes           1.72    26
6 Greedo                   74    173 <NA>       yes           1.73    25
7 Anakin Skywalker         84    188 blond      yes           1.88    24
8 Finis Valorum            NA    170 blond      yes           1.7     NA
```

Mutate (add or modify columns)

```
ds <- starwars %>% select(name, mass, height, hair_color)
ds <- ds %>% mutate(in_movie = "yes")
ds <- ds %>% mutate(height_m = height/100,
                    bmi = mass/(height_m^2),
                    bmi = round(bmi)) %>%
            arrange(desc(bmi))

ds <- ds %>%
  filter(hair_color %in% c("blond", NA)) %>%
  mutate(hair_color = ifelse(is.na(hair_color),"no hair", hair_color),
         mass = ifelse(mass > 1000, "huge", "not huge"))

ds
```

```
# A tibble: 8 × 7
  name                 mass     height hair_color in_movie height_m   bmi
  <chr>                <chr>     <int> <chr>      <chr>       <dbl> <dbl>
1 Jabba Desilijic Tiure huge        175 no hair    yes          1.75   443
2 R2-D2                not huge      96 no hair    yes          0.96    35
3 R5-D4                not huge      97 no hair    yes          0.97    34
4 C-3PO                not huge     167 no hair    yes          1.67    27
5 Luke Skywalker       not huge     172 blond      yes          1.72    26
6 Greedo               not huge     173 no hair    yes          1.73    25
7 Anakin Skywalker     not huge     188 blond      yes          1.88    24
8 Finis Valorum        <NA>         170 blond      yes          1.7     NA
```

Tricky, common task: Changing some but not all values within a column

```
# Change all NA values to "no hair", keep all others the same
ds <- ds %>% mutate(hair_color = ifelse(is.na(hair_color),"no hair", hair_color))

# Set all heights greater than 100 to 100, keep all others the same
ds <- ds %>% mutate(height = ifelse(height > 100,100, height))
```

Summarize (calculate variables in aggregate)

```
ds <- starwars %>% select(name, mass, height, species)
```

Summarize (calculate variables in aggregate)

```
ds <- starwars %>% select(name, mass, height, species)
ds %>% summarize(min_height = min(height))
```

```
# A tibble: 1 × 1
  min_height
       <int>
1         NA
```

Summarize (calculate variables in aggregate)

```
ds <- starwars %>% select(name, mass, height, species)
ds %>% summarize(min_height = min(height))
ds %>% summarize(min_height = min(height, na.rm = T))
```

```
# A tibble: 1 × 1
  min_height
       <int>
1         NA

# A tibble: 1 × 1
  min_height
       <int>
1         66
```

Summarize (calculate variables in aggregate)

```
ds <- starwars %>% select(name, mass, height, species)
ds %>% summarize(min_height = min(height))
ds %>% summarize(min_height = min(height, na.rm = T))

ds %>% summarize(min_height = min(height, na.rm = T),
                 m_height = mean(height, na.rm = T),
                 max_height = max(height, na.rm = T))
```

```
# A tibble: 1 × 1
  min_height
       <int>
1         NA

# A tibble: 1 × 1
  min_height
       <int>
1         66

# A tibble: 1 × 3
  min_height m_height max_height
       <int>    <dbl>      <int>
1         66     174.        264
```

Summarize (calculate variables in aggregate)

```
ds <- starwars %>% select(name, mass, height, species)
ds %>% summarize(min_height = min(height))
ds %>% summarize(min_height = min(height, na.rm = T))

ds %>% summarize(min_height = min(height, na.rm = T),
                 m_height = mean(height, na.rm = T),
                 max_height = max(height, na.rm = T))

ds %>% group_by(species)
```

```
# A tibble: 1 × 1
  min_height
       <int>
1         NA

# A tibble: 1 × 1
  min_height
       <int>
1         66

# A tibble: 1 × 3
  min_height m_height max_height
       <int>    <dbl>      <int>
1         66     174.        264

# A tibble: 87 × 4
# Groups:   species [38]
   name                mass height species
   <chr>              <dbl>  <int> <chr>
 1 Luke Skywalker        77    172 Human
 2 C-3PO                 75    167 Droid
 3 R2-D2                 32     96 Droid
 4 Darth Vader          136    202 Human
 5 Leia Organa           49    150 Human
 6 Owen Lars            120    178 Human
```

Summarize (calculate variables in aggregate)

```
ds <- starwars %>% select(name, mass, height, species)
ds %>% summarize(min_height = min(height))
ds %>% summarize(min_height = min(height, na.rm = T))

ds %>% summarize(min_height = min(height, na.rm = T),
                 m_height = mean(height, na.rm = T),
                 max_height = max(height, na.rm = T))

ds %>% group_by(species) %>%
  summarize(min_height = min(height, na.rm = T),
                 m_height = mean(height, na.rm = T),
                 max_height = max(height, na.rm = T),
                 n = n())
```

```
# A tibble: 1 × 1
  min_height
       <int>
1         NA

# A tibble: 1 × 1
  min_height
       <int>
1         66

# A tibble: 1 × 3
  min_height m_height max_height
       <int>    <dbl>      <int>
1         66     174.        264

# A tibble: 38 × 5
   species   min_height m_height max_height     n
   <chr>          <int>    <dbl>      <int> <int>
 1 Aleena            79     79           79     1
 2 Besalisk         198    198          198     1
 3 Cerean           198    198          198     1
 4 Chagrian         196    196          196     1
 5 Clawdite         168    168          168     1
 6 Droid             96    131.         200     6
 7 Dug              112    112          112     1
```