Data wrangling fundamentals

What do you do with data?

- Data manipulation and cleaning
- Calculate summary statistics
- Visualization
- Input for modeling

Data manipulation

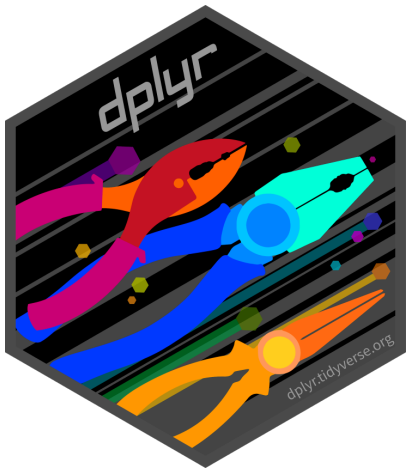
```
1 glimpse(starwars)
```

```
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",
```

What manipulation might I want to do with the `starwars` data?

- might handle NAs ...

`dplyr`: Tools for data wrangling



- part of the tidyverse
- provides a “grammar of data manipulation”: useful verbs (functions) for manipulating data
- we will cover the key `dplyr` functions

Some core verbs for data wrangling

- `filter()`: take a subset of the rows (i.e., observations)
- `select()`: take a subset of the columns (i.e., features, variables)
- `mutate()`: add or modify existing columns
- `arrange()`: sort the rows
- `group_by()`: group rows by one or more variables
- `summarize()`: aggregate the data across rows (often after grouping)

Creating a subset of the rows

Question: Suppose I only want the droids in the `starwars` data. How would I choose only those rows?

Creating a subset of the rows

Question: Suppose I only want the droids in the `starwars` data. How would I choose only those rows?

```
1 filter(starwars, species == "Droid")
```

```
# A tibble: 6 × 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
masculi…
2 R2-D2      96    32 <NA>       white, blue red                33 none
masculi…
3 R5-D4      97    32 <NA>       white, red  red                NA none
masculi…
4 IG-88     200   140 none       metal       red                15 none
masculi…
```

Creating a subset of the rows

```r
starwars |>
  filter(species == "Droid")
```

```
# 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 C-3PO    167    75 <NA>       gold        yellow            112 none
masculine
2 R2-D2     96    32 <NA>       white, blue red                33 none
masculine
# ℹ 5 more variables: homeworld <chr>, species <chr>, films <list>,
#   vehicles <list>, starships <list>
```

pipe

|> means "take this, then do that"

Calculating summary statistics

Question: What is the average height for droids in the dataset?

- focus on droids
- calculate summary statistics

Calculating summary statistics

Question: What is the average height for droids in the dataset?

```
1 starwars |>
2   filter(species == "Droid") |>
3   summarize(mean_height = mean(height))
```

focusing on droids

calculate a summary statistic

```
# A tibble: 1 × 1
  mean_height
        <dbl>
1          NA
```

- pipes (`|>`) can be chained together
- `summarize` calculates summary statistics
- Why am I getting `NA`?

Handling missing values

```
# A tibble: 6 × 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
masculi…
2 R2-D2      96    32 <NA>       white, blue red                33 none
masculi…
3 R5-D4      97    32 <NA>       white, red  red                NA none
masculi…
4 IG-88     200   140 none       metal       red                15 none
masculi…
```

```
1 starwars |>
2   filter(species == "Droid") |>
3   summarize(mean_height = mean(height, na.rm=T))
```

"ignore missing values (NAs) when calculating mean"

calculate mean

```
# A tibble: 1 × 1
  mean_height
        <dbl>
1        131.
```

Calculating summary statistics

Question: What if I want the average height for *humans*?

```
1 starwars |>
2   filter(species == "Droid") |>
3   summarize(mean_height = mean(height, na.rm=T))
```

Calculating summary statistics

Question: What if I want the average height for *humans*?

```
starwars |>
  filter(species == "Human") |>
  summarize(mean_height = mean(height, na.rm=T))
```

```
# A tibble: 1 × 1
  mean_height
        <dbl>
1         178
```

Calculating summary statistics

Question: What is the average height for *each* species?

Calculating summary statistics

Question: What is the average height for *each* species?

```
1 starwars |>
2   group_by(species) |>
3   summarize(mean_height = mean(height, na.rm=T))
```

group by species

calculate summary statistics within each group

```
# A tibble: 38 × 2
   species   mean_height
   <chr>           <dbl>
 1 Aleena           79
 2 Besalisk        198
 3 Cerean          198
 4 Chagrian        196
 5 Clawdite        168
 6 Droid           131.
 7 Dug             112
 8 Ewok             88
 9 Geonosian       183
10 Gungan          209.
```

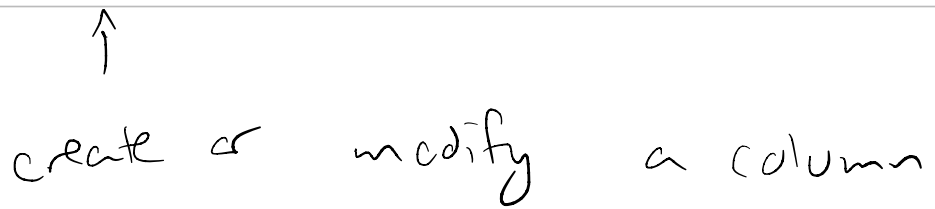
Creating new variables

Question: What is the distribution of the ratio of body mass to height?

Creating new variables

Question: What is the distribution of the ratio of body mass to height?

```
1 starwars |>
2   mutate(body_ratio = mass/height)
```



Creating new variables

```
1 starwars |>
2   mutate(body_ratio = mass/height) |>
3   group_by(species) |>
4   summarize(mean_ratio = mean(body_ratio, na.rm=T),
5             sd_ratio = sd(body_ratio, na.rm=T))
```

```
# A tibble: 38 × 3
   species   mean_ratio sd_ratio
   <chr>          <dbl>    <dbl>
 1 Aleena         0.190  NA
 2 Besalisk       0.515  NA
 3 Cerean         0.414  NA
 4 Chagrian     NaN      NA
 5 Clawdite       0.327  NA
 6 Droid          0.453   0.174
 7 Dug            0.357  NA
 8 Ewok           0.227  NA
 9 Geonosian      0.437  NA
10 Gungan         0.351   0.0207
```

Creating new variables

```
1 starwars |>
2   mutate(body_ratio = mass/height) |>
3   group_by(species) |>
4   summarize(mean_ratio = mean(body_ratio, na.rm=T),
5             sd_ratio = sd(body_ratio, na.rm=T),
6             N = n())
```

counts # rows

```
# A tibble: 38 × 4
   species   mean_ratio sd_ratio     N
   <chr>          <dbl>    <dbl> <int>
 1 Aleena         0.190  NA           1
 2 Besalisk       0.515  NA           1
 3 Cerean         0.414  NA           1
 4 Chagrian     NaN      NA           1
 5 Clawdite       0.327  NA           1
 6 Droid          0.453   0.174       6
 7 Dug            0.357  NA           1
 8 Ewok           0.227  NA           1
 9 Geonosian      0.437  NA           1
10 Gungan         0.351   0.0207      3
```

Summary so far

- `filter`: choose certain rows
- `summarize`: calculate summary statistics
- `group_by`: group rows together
- `mutate`: create new columns

Class activity

https://sta279-f25.github.io/class_activities/ca_02.html

- Work with a neighbor on the class activity
- At the end of class, submit your work as an HTML file on Canvas (one per group, list all your names)
- I will come around and answer any questions

For next time, read:

- Chapter 3 in *R for Data Science* (2nd ed.)
- Chapter 4 in *Modern Data Science with R*