Online CDS 101 *Introduction to Computational and Data Sciences*

Data visualization

Data visualization with ggplot2

Star Wars data

Loading `tidyverse` also loads a dataset called `starwars` into your RStudio environment:

```r
library(tidyverse)
starwars
```

```
## # A tibble: 87 x 13
##    name  height  mass hair_color skin_color eye_color birth_year gender
##    <chr>  <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr>
##  1 Luke…    172    77 blond      fair       blue            19   male
##  2 C-3PO    167    75 <NA>       gold       yellow         112   <NA>
##  3 R2-D2     96    32 <NA>       white, bl… red             33   <NA>
##  4 Dart…    202   136 none       white      yellow          41.9 male
##  5 Leia…    150    49 brown      light      brown           19   female
##  6 Owen…    178   120 brown, gr… light      blue            52   male
##  7 Beru…    165    75 brown      light      blue            47   female
##  8 R5-D4     97    32 <NA>       white, red red             NA   <NA>
##  9 Bigg…    183    84 black      light      brown           24   male
## 10 Obi-…    182    77 auburn, w… fair       blue-gray       57   male
## # ... with 77 more rows, and 5 more variables: homeworld <chr>,
## #   species <chr>, films <list>, vehicles <list>, starships <list>
```

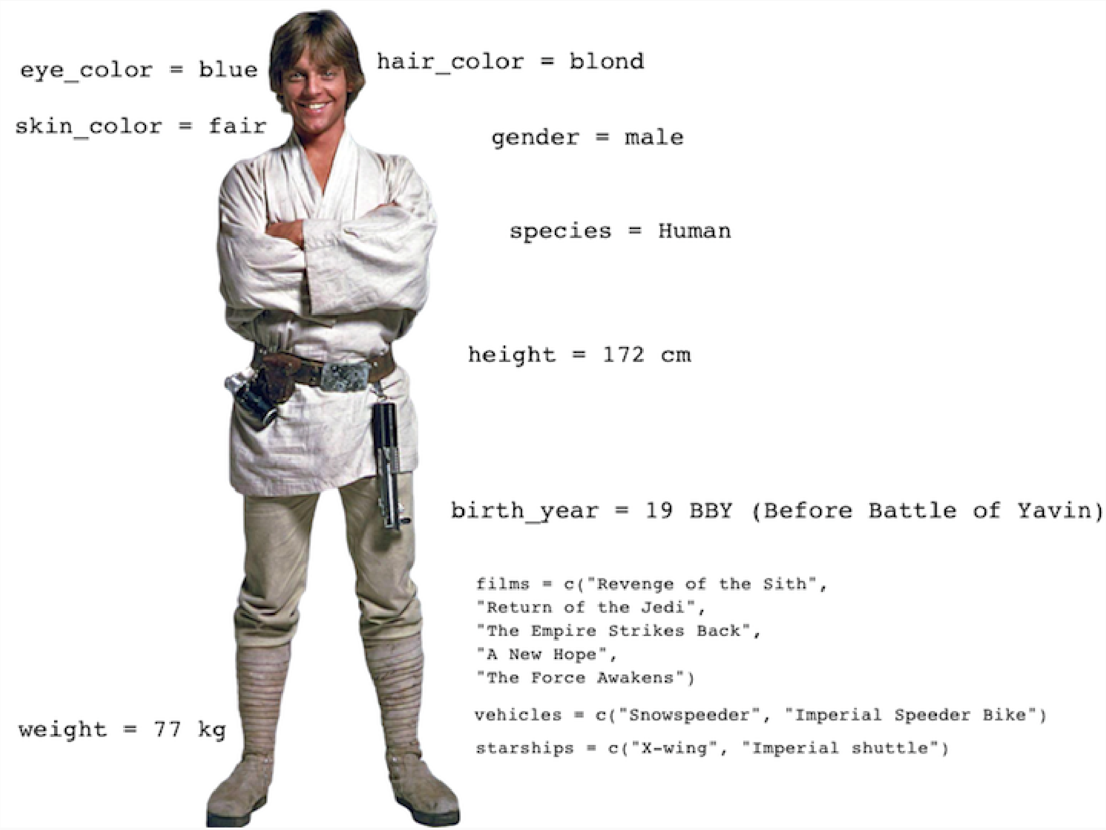
Dataset terminology

What does each row represent? What does each column represent?

```
starwars
```

```
## # A tibble: 87 x 13
##    name  height  mass hair_color skin_color eye_color birth_year gender
##    <chr>  <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr>
##  1 Luke…    172    77 blond      fair       blue            19   male
##  2 C-3PO    167    75 <NA>       gold       yellow         112   <NA>
##  3 R2-D2     96    32 <NA>       white, bl… red             33   <NA>
##  4 Dart…    202   136 none       white      yellow          41.9 male
##  5 Leia…    150    49 brown      light      brown           19   female
##  6 Owen…    178   120 brown, gr… light      blue            52   male
##  7 Beru…    165    75 brown      light      blue            47   female
##  8 R5-D4     97    32 <NA>       white, red red             NA   <NA>
##  9 Bigg…    183    84 black      light      brown           24   male
## 10 Obi-…    182    77 auburn, w… fair       blue-gray       57   male
## # ... with 77 more rows, and 5 more variables: homeworld <chr>,
## #   species <chr>, films <list>, vehicles <list>, starships <list>
```

Luke Skywalker



eye_color = blue

skin_color = fair

hair_color = blond

gender = male

species = Human

height = 172 cm

birth_year = 19 BBY (Before Battle of Yavin)

```
films = c("Revenge of the Sith",
"Return of the Jedi",
"The Empire Strikes Back",
"A New Hope",
"The Force Awakens")
vehicles = c("Snowspeeder", "Imperial Speeder Bike")
starships = c("X-wing", "Imperial shuttle")
```

weight = 77 kg

What's in the Star Wars data?

Take a `glimpse` at the data:

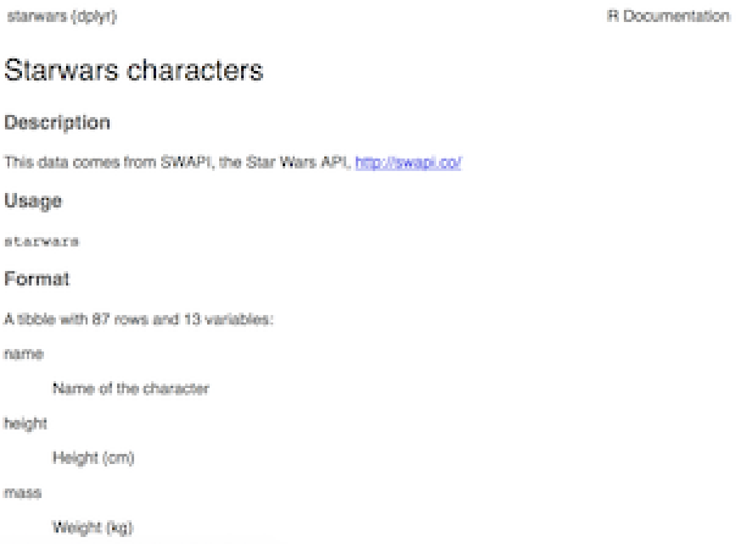
```
glimpse(starwars)
```

```
## Observations: 87
## Variables: 13
## $ name       <chr> "Luke Skywalker", "C-3PO", "R2-D2", "Darth Vader", ...
## $ height     <int> 172, 167, 96, 202, 150, 178, 165, 97, 183, 182, 188...
## $ mass       <dbl> 77.0, 75.0, 32.0, 136.0, 49.0, 120.0, 75.0, 32.0, 8...
## $ hair_color <chr> "blond", NA, NA, "none", "brown", "brown, grey", "b...
## $ skin_color <chr> "fair", "gold", "white, blue", "white", "light", "l...
## $ eye_color  <chr> "blue", "yellow", "red", "yellow", "brown", "blue",...
## $ birth_year <dbl> 19.0, 112.0, 33.0, 41.9, 19.0, 52.0, 47.0, NA, 24.0...
## $ gender     <chr> "male", NA, NA, "male", "female", "male", "female",...
## $ homeworld  <chr> "Tatooine", "Tatooine", "Naboo", "Tatooine", "Alder...
## $ species    <chr> "Human", "Droid", "Droid", "Human", "Human", "Human...
## $ films      <list> [<"Revenge of the Sith", "Return of the Jedi", "Th...
## $ vehicles   <list> [<"Snowspeeder", "Imperial Speeder Bike">, <>, <>,...
## $ starships  <list> [<"X-wing", "Imperial shuttle">, <>, <>, "TIE Adva...
```

What's in the Star Wars data?

Run the following **in the Console** to view the help

```
?starwars
```



starwars {dplyr} R Documentation

Starwars characters

Description

This data comes from SWAPI, the Star Wars API, http://swapi.co/

Usage

```
starwars
```

Format

A tibble with 87 rows and 13 variables:

name

Name of the character

height

Height (cm)

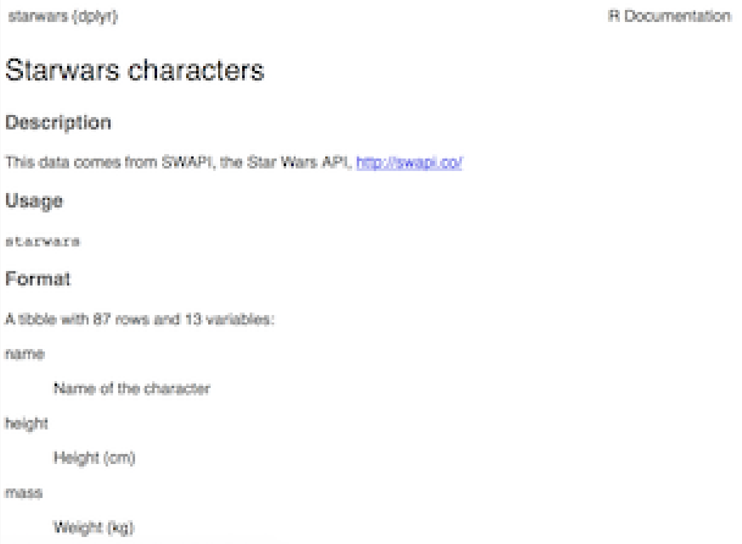
mass

Weight (kg)

What's in the Star Wars data?

Run the following **in the Console** to view the help

```
?starwars
```



starwars {dplyr} R Documentation

Starwars characters

Description

This data comes from SWAPI, the Star Wars API, http://swapi.co/

Usage

```
starwars
```

Format

A tibble with 87 rows and 13 variables:

name

Name of the character

height

Height (cm)

mass

Weight (kg)

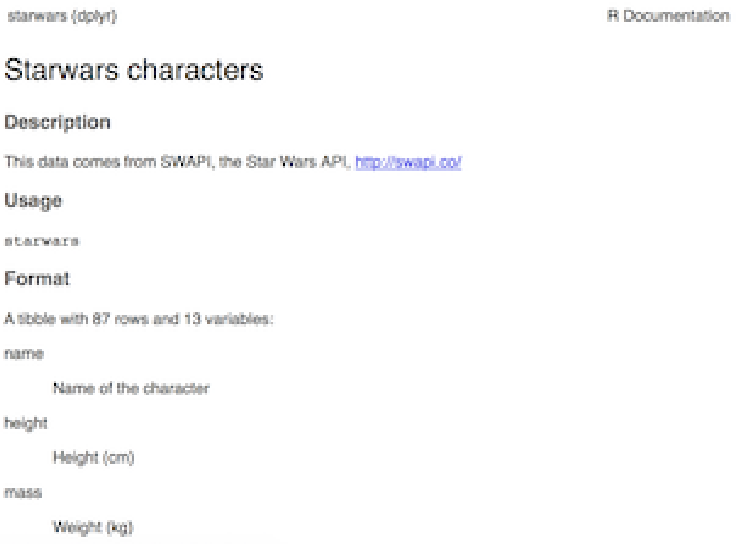
How many rows and columns does this dataset have?

What does each row represent? What does each column represent?

What's in the Star Wars data?

Run the following **in the Console** to view the help

```
?starwars
```



starwars {dplyr}   R Documentation

Starwars characters

Description

This data comes from SWAPI, the Star Wars API, http://swapi.co/

Usage

```
starwars
```

Format

A tibble with 87 rows and 13 variables:

name

    Name of the character

height

    Height (cm)

mass

    Weight (kg)

How many rows and columns does this dataset have?

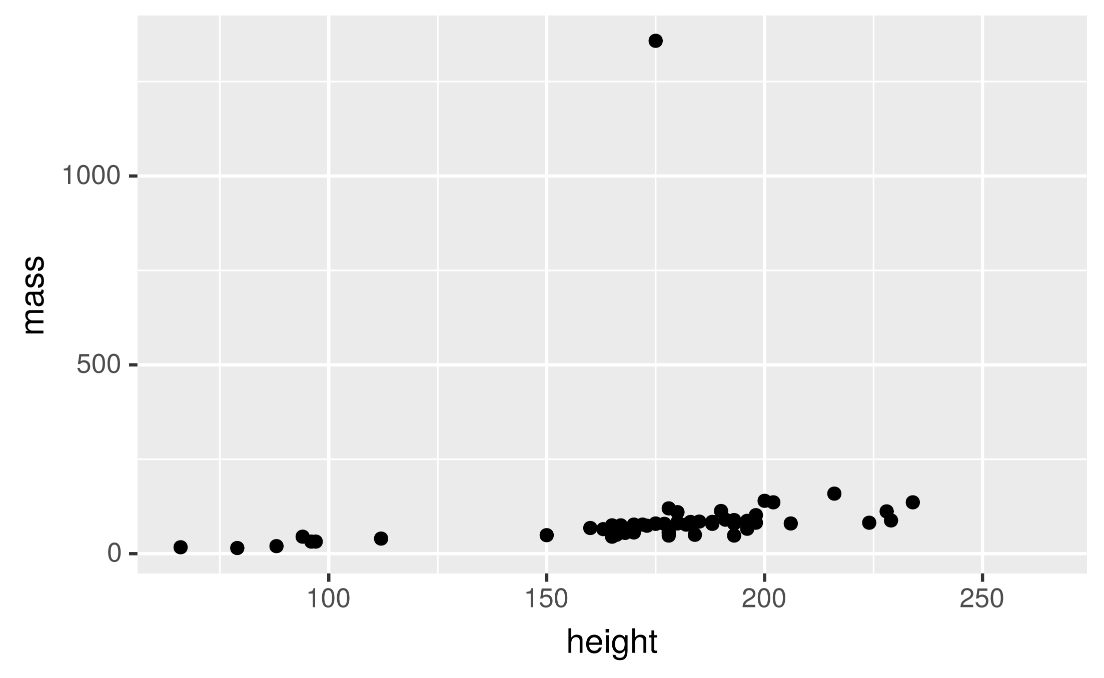
What does each row represent? What does each column represent?

Make a prediction: What relationship do you expect to see between height and mass?

Mass vs. height (`geom_point()`)

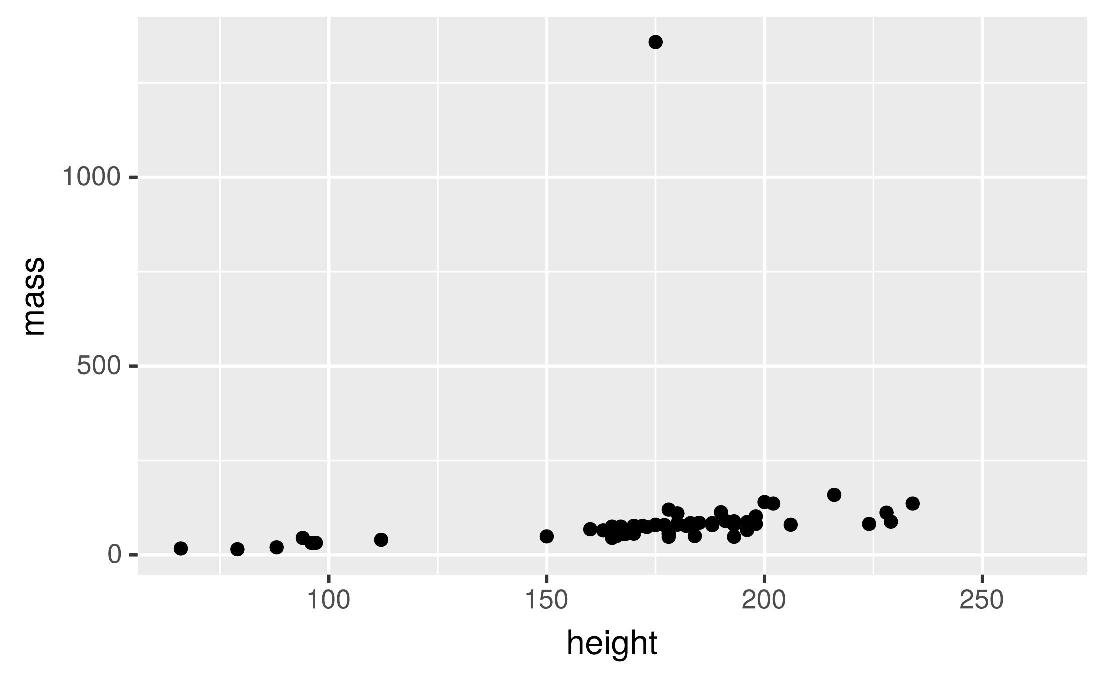
Not all characters have height and mass information (hence 28 of them not plotted)

```
ggplot(data = starwars) +
  geom_point(mapping = aes(x = height, y = mass))
```



Mass vs. height

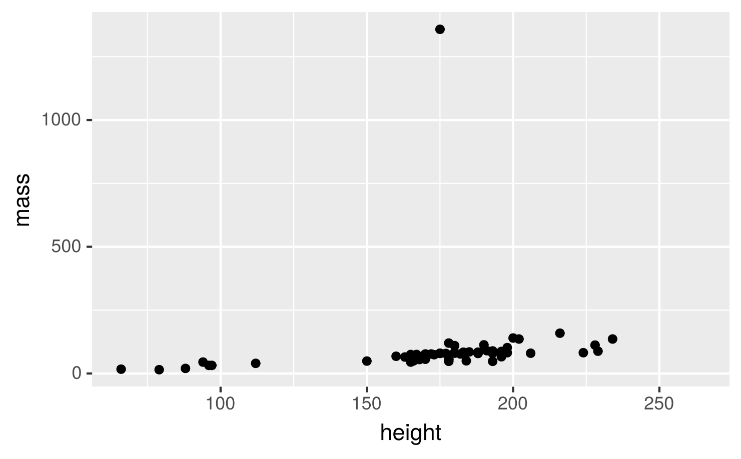
How would you describe this relationship? What other variables would help us understand data points that don't follow the overall trend?



Mass vs. height

Who is the not so tall but really massive character?

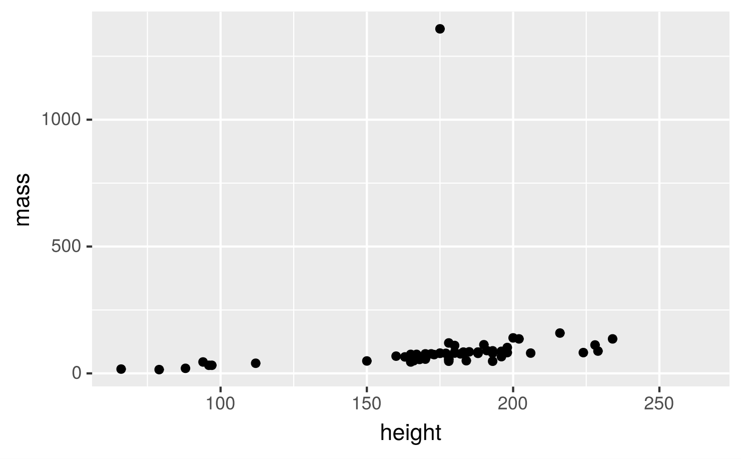
```
ggplot(data = starwars) +
  geom_point(mapping = aes(x = height, y = mass))
```



Mass vs. height

Who is the not so tall but really massive character?

```
ggplot(data = starwars) +
  geom_point(mapping = aes(x = height, y = mass))
```



Additional variables

Can display additional variables with

- aesthetics (like shape, colour, size), or
- faceting (small multiples displaying different subsets)

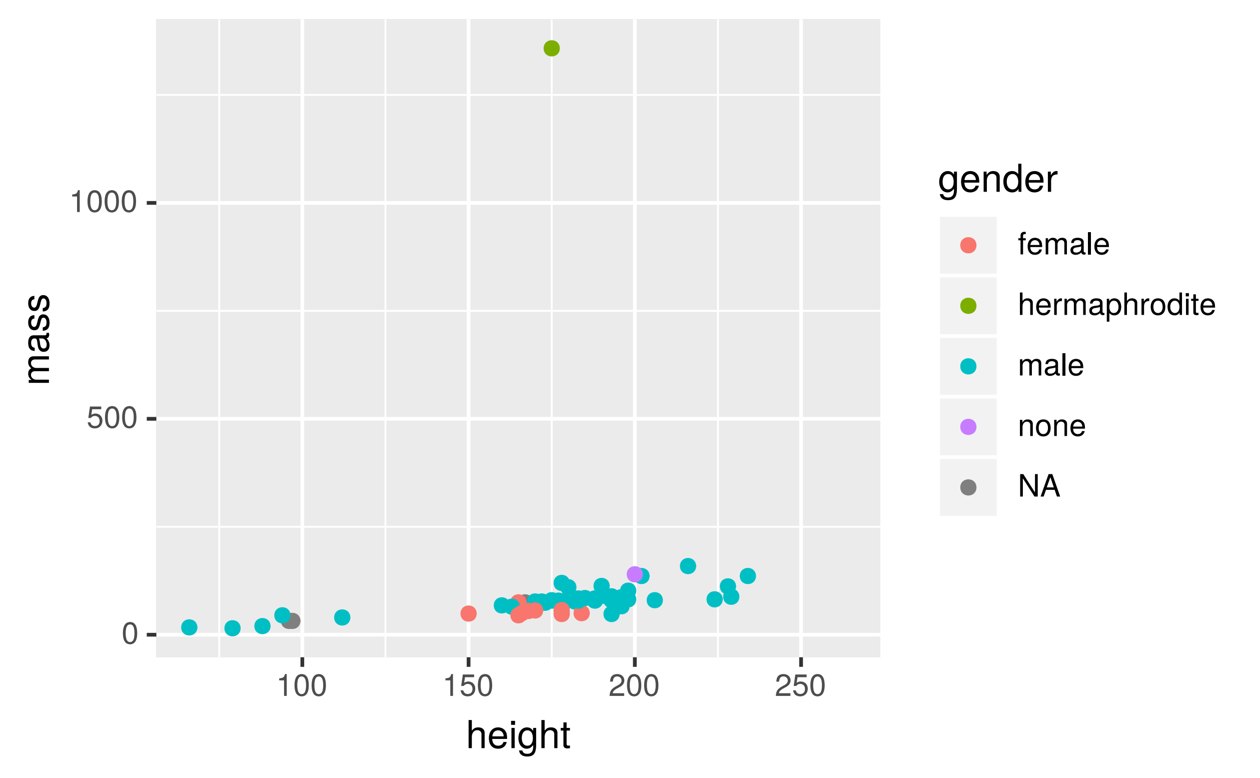
Aesthetics

Visual characteristics of plotting characters that can be **mapped to data** are

- `color`
- `size`
- `shape`
- `alpha` (transparency)

Mass vs. height + gender

```
ggplot(data = starwars) +
  geom_point(mapping = aes(x = height, y = mass, color = gender))
```



Aesthetics summary

- Continuous variable are measured on a continuous scale
- Discrete variables are measured (or often counted) on a discrete scale

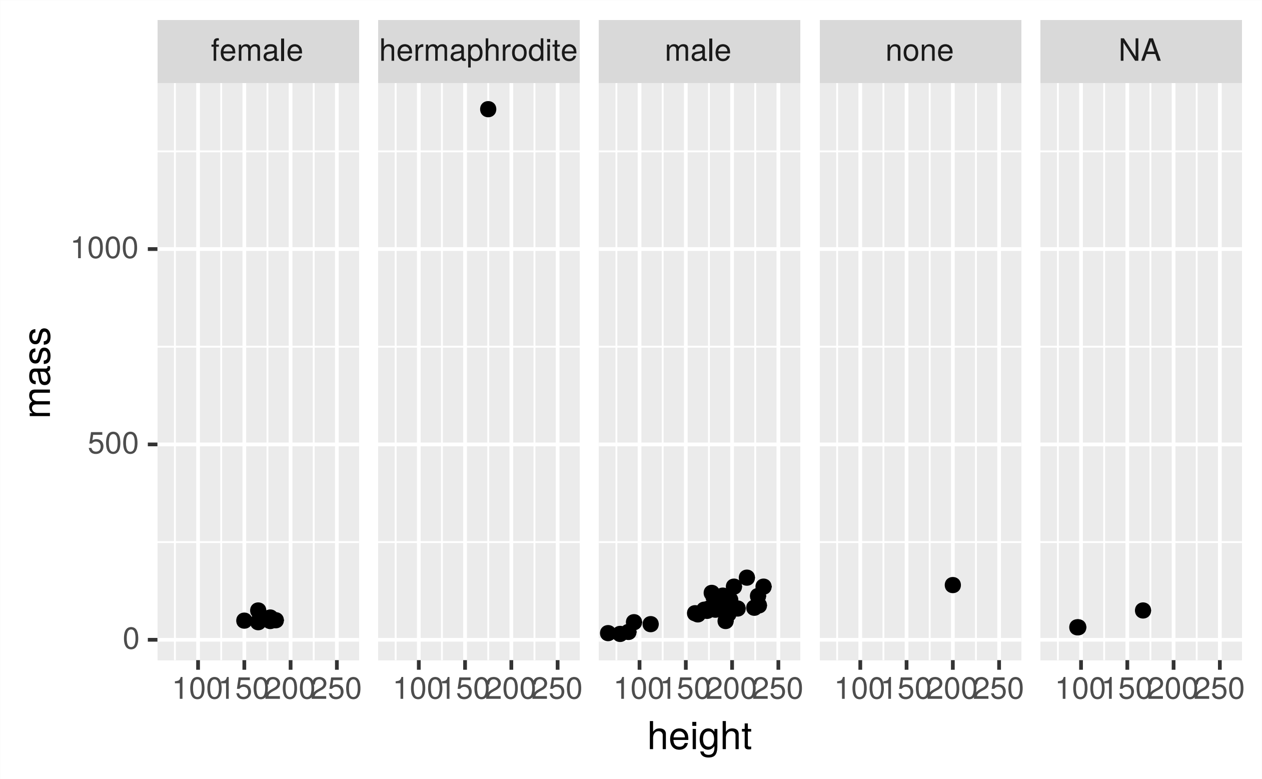
aesthetics	discrete	continuous
color	rainbow of colors	gradient
size	discrete steps	linear mapping between radius and value
shape	different shape for each	shouldn't (and doesn't) work

Faceting

- Smaller plots that display different subsets of the data
- Useful for exploring conditional relationships and large data

Mass vs. height by gender

```
ggplot(data = starwars) +
  geom_point(mapping = aes(x = height, y = mass)) +
  facet_grid(. ~ gender)
```

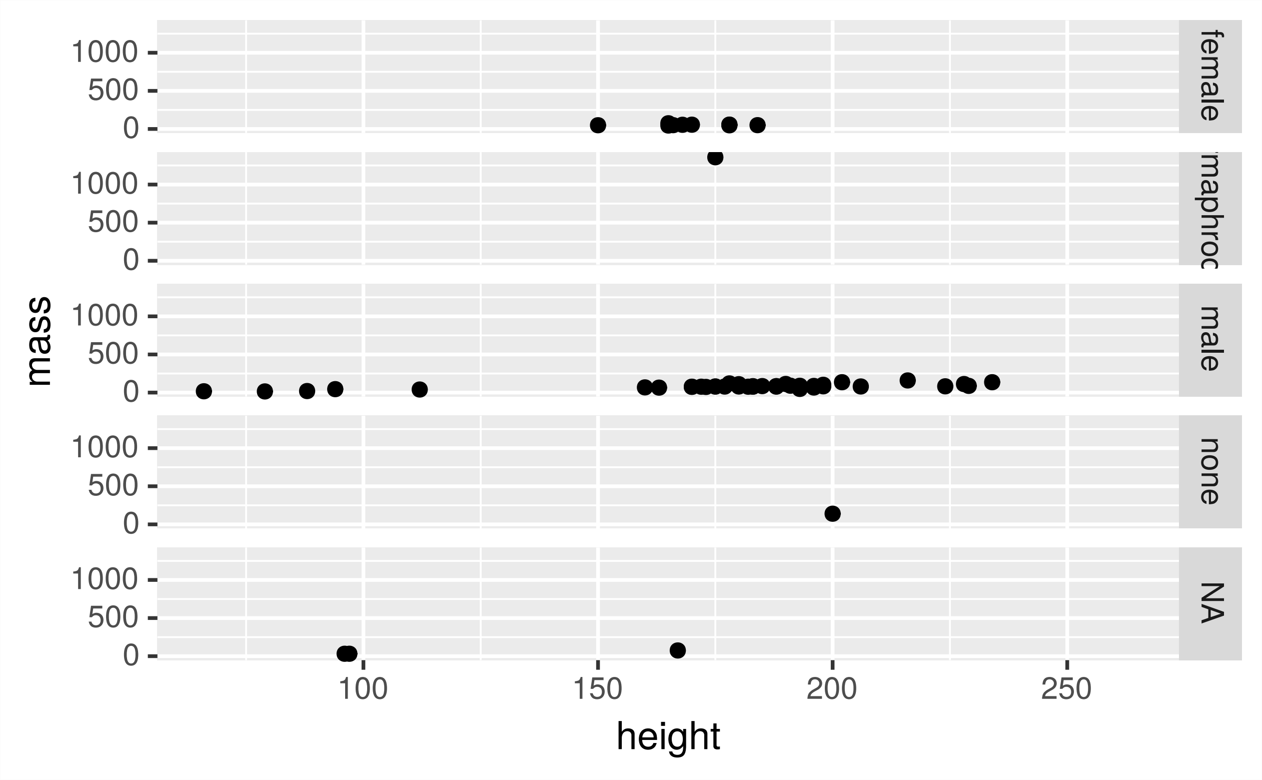


Many ways to facet

In the next few examples, think about what each plot displays. Think about how the code relates to the output.

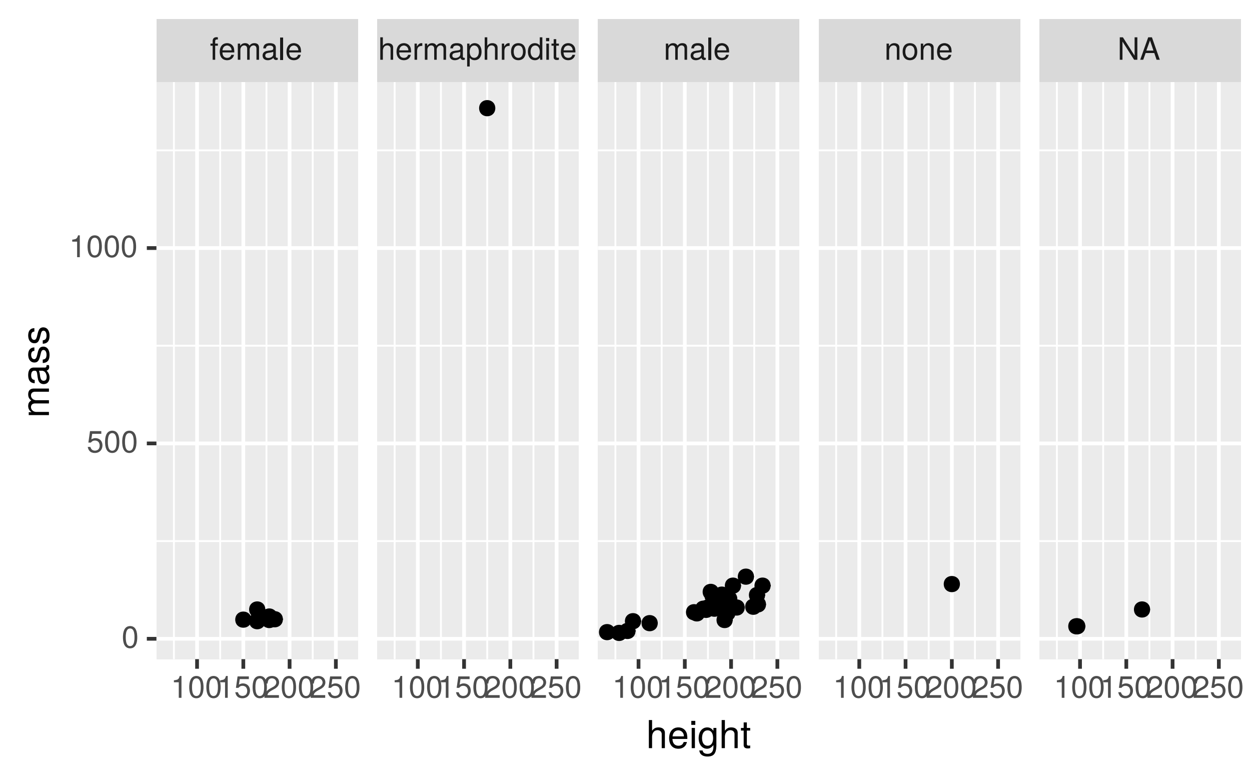
Many ways to facet

```
ggplot(data = starwars) +
  geom_point(mapping = aes(x = height, y = mass)) +
  facet_grid(gender ~ .)
```



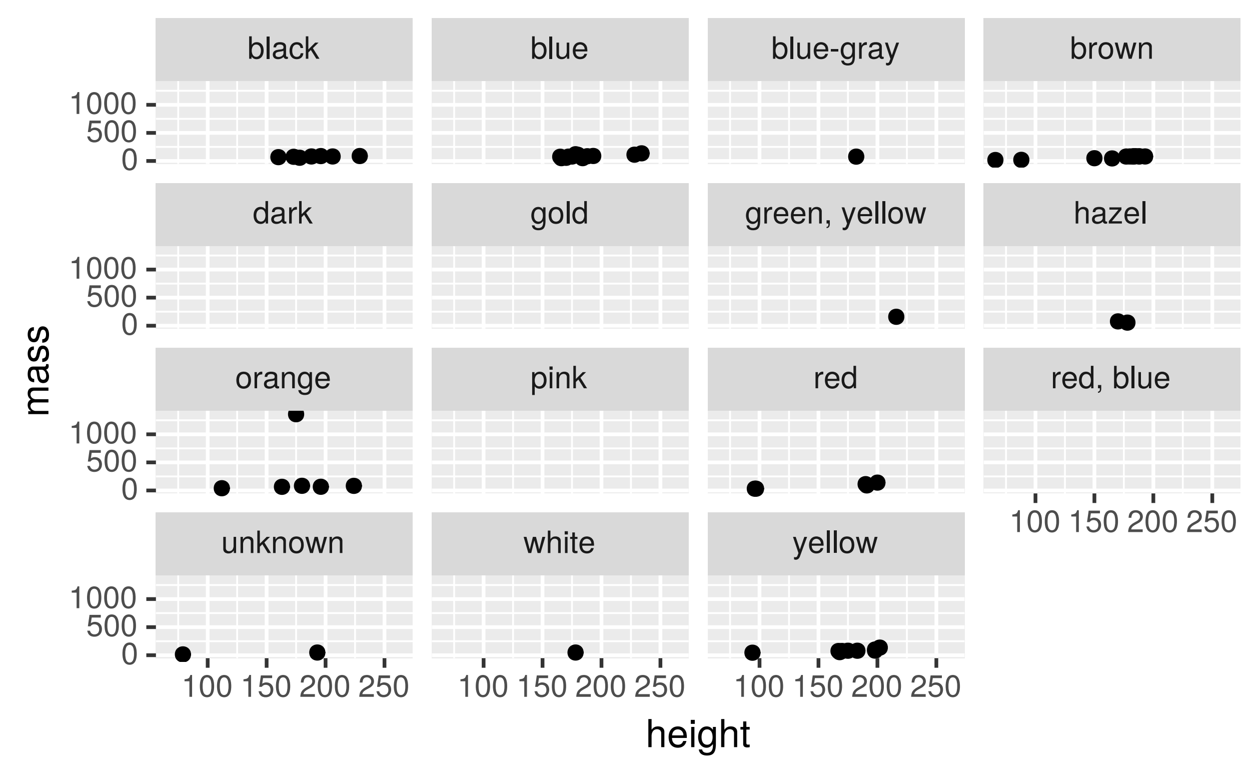
Many ways to facet

```
ggplot(data = starwars) +
  geom_point(mapping = aes(x = height, y = mass)) +
  facet_grid(. ~ gender)
```



Many ways to facet

```
ggplot(data = starwars) +
  geom_point(mapping = aes(x = height, y = mass)) +
  facet_wrap(~ eye_color)
```



Facet summary

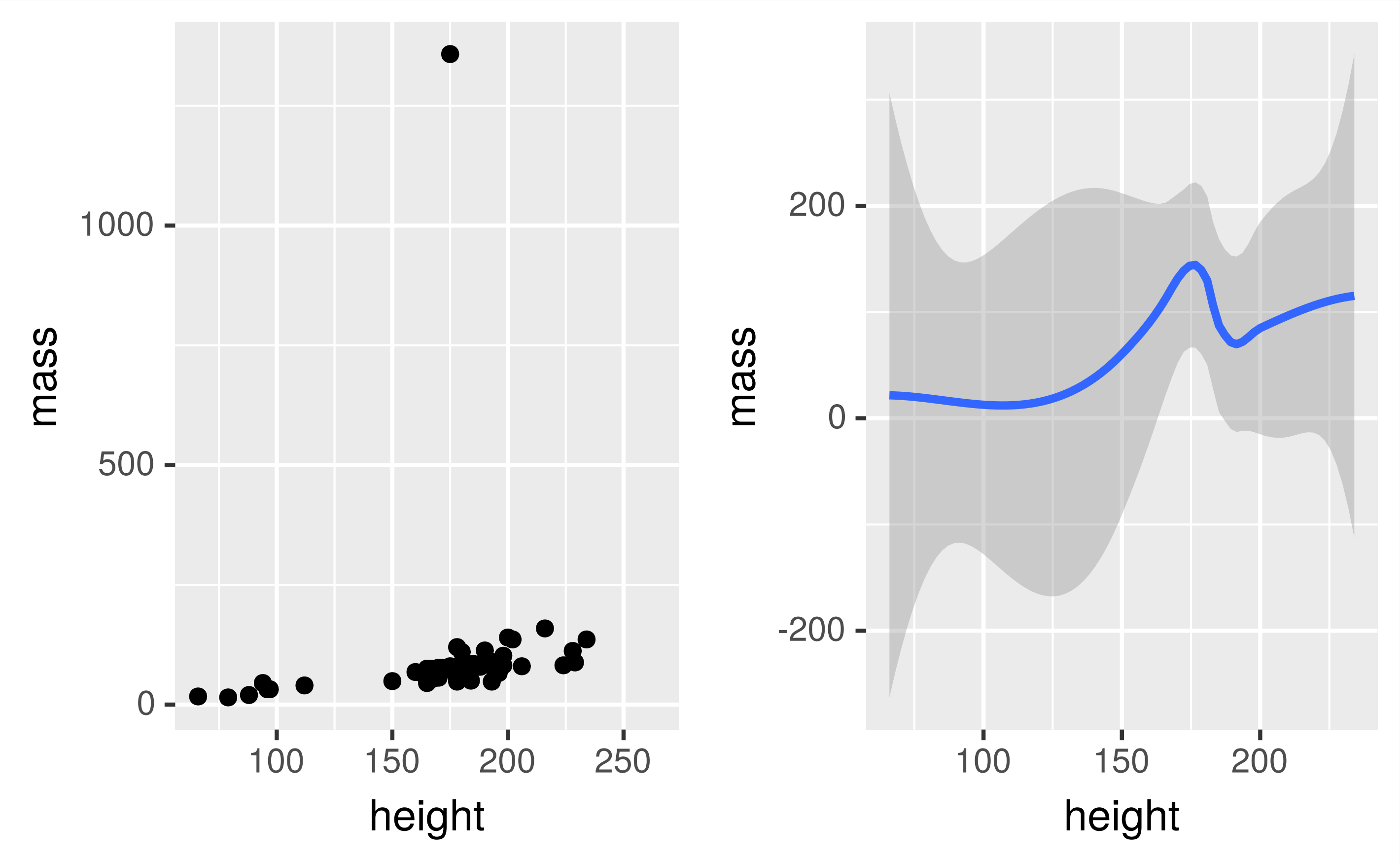
- `facet_grid()`: 2d grid, rows ~ cols, . for no split
- `facet_wrap()`: 1d ribbon wrapped into 2d

Other geoms

How are these plots similar? How are they different?

Other geoms

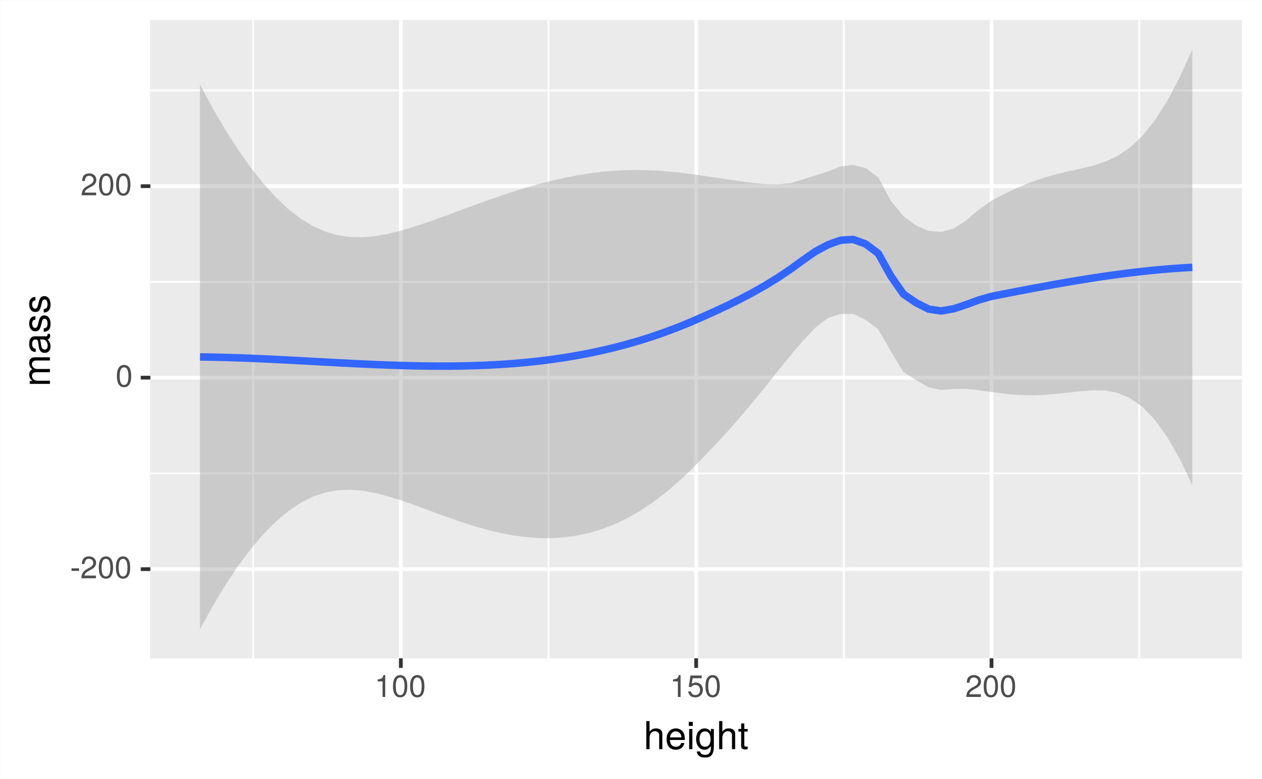
How are these plots similar? How are they different?



geom_smooth

To plot a smooth curve, use `geom_smooth()`

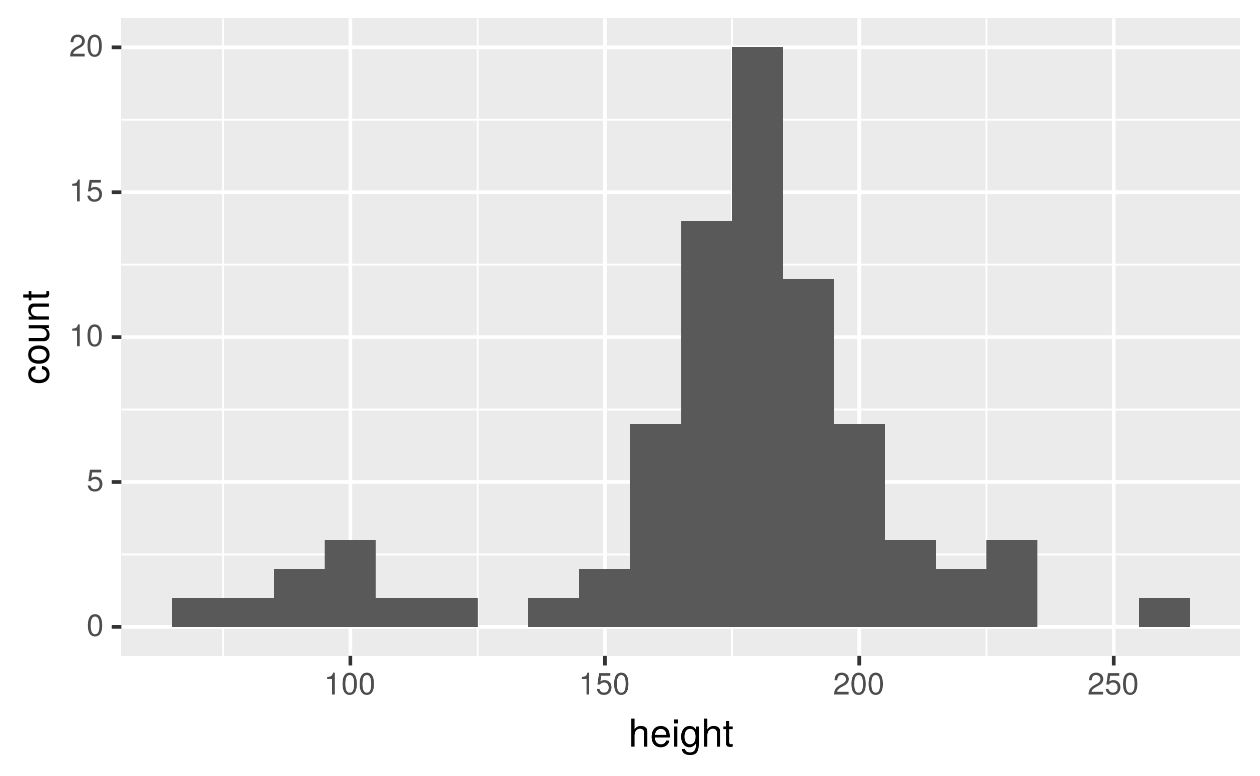
```
ggplot(data = starwars) +
  geom_smooth(mapping = aes(x = height, y = mass))
```



Histograms

For numerical variables

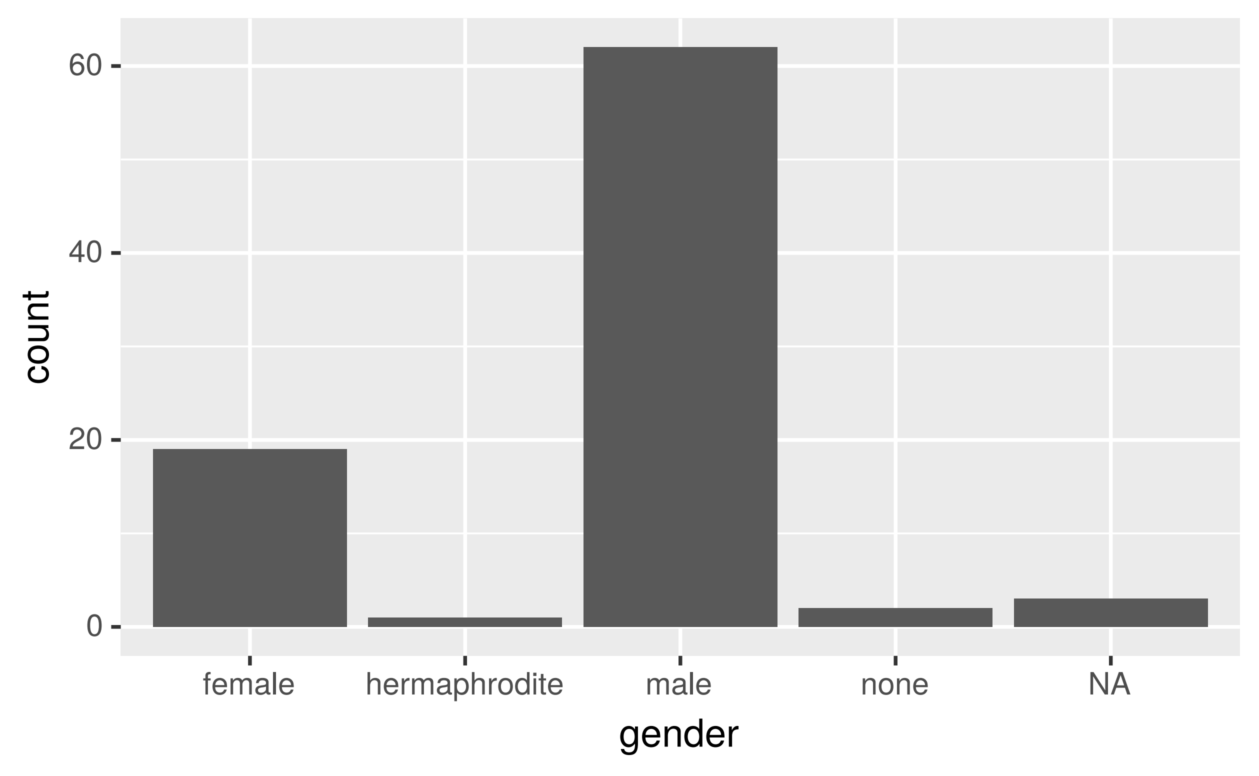
```
ggplot(starwars) +
  geom_histogram(mapping = aes(x = height), binwidth = 10)
```



Bar plots

For categorical variables

```
ggplot(starwars) +
  geom_bar(mapping = aes(x = gender))
```



Credits

License

Acknowledgments

Content adapted from the Fundamentals of data & data visualization slides developed by Mine Çetinkaya-Rundel and made available under the CC BY license.