Section 001
CDS 101 *Introduction to Computational and Data Sciences*

Class 7: Data visualization III

February 13, 2018

Annoucements

- No reading for Thursday's class
- Come talk to me ASAP if...
  - ...you are still experiencing issues with using Github to submit assignments
  - ...your RStudio installation continues to give you unknown errors
- Website will be updated soon with prior lecture's slides and homework 1

Data visualization with `ggplot2`

Structure of R commands

Functions in R are often verbs, and then in parantheses are the arguments for those functions.

```
verb(what-you-want-to-apply-verb-to, other-arguments)
```

For example:

Structure of R commands

Functions in R are often verbs, and then in parantheses are the arguments for those functions.

```
verb(what-you-want-to-apply-verb-to, other-arguments)
```

For example:

```
glimpse(mpg)            # Glimpse into the mpg dataset
```

Structure of R commands

Functions in R are often verbs, and then in parantheses are the arguments for those functions.

```
verb(what-you-want-to-apply-verb-to, other-arguments)
```

For example:

```
glimpse(mpg)              # Glimpse into the mpg dataset
```

```
ggplot(mpg) +                                # Create plot window; plot
                                             #     variables found in mpg
                                             #     dataset
  geom_point(aes(x = displ, y = hwy))        # Create scatterplot with displ
                                             #     variable on x-axis, hwy
                                             #     variable on y-axis
```

Structure of ggplot2 commands

To use ggplot2 functions, load `tidyverse`:

```
library(tidyverse)
```

Structure of ggplot2 commands

To use ggplot2 functions, load `tidyverse`:

```
library(tidyverse)
```

In ggplot2 the structure of the code for plots can often be summarized as

```
ggplot +
  geom_word
```

Structure of ggplot2 commands

To use ggplot2 functions, load `tidyverse`:

```
library(tidyverse)
```

In ggplot2 the structure of the code for plots can often be summarized as

```
ggplot +
  geom_word
```

or, more precisely

Structure of ggplot2 commands

To use ggplot2 functions, load `tidyverse`:

```
library(tidyverse)
```

In ggplot2 the structure of the code for plots can often be summarized as

```
ggplot +
  geom_word
```

or, more precisely

```
ggplot(data = [dataset]) +
   geom_word(mapping = aes(x = [x-variable], y = [y-variable])) +
   other options
```

Structure of `ggplot2` commands

To use ggplot2 functions, load `tidyverse`:

```
library(tidyverse)
```

In ggplot2 the structure of the code for plots can often be summarized as

```
ggplot +
  geom_word
```

or, more precisely

```
ggplot(data = [dataset]) +
   geom_word(mapping = aes(x = [x-variable], y = [y-variable])) +
   other options
```

`Geoms`, short for geometric objects, describe the type of plot you will produce.

About ggplot2

- ggplot2 is the name of the package
- The `gg` in "ggplot2" stands for Grammar of Graphics
- Inspired by the book **Grammar of Graphics** by Lee Wilkinson
- `ggplot()` is the main function in ggplot2

Visualizing Star Wars

Star Wars data

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

```
library(tidyverse)
starwars
```

```
## # A tibble: 87 x 13
##                  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, and 7 more variables: birth_year <dbl>,
## #   gender <chr>, homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

Dataset terminology

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

```
## # A tibble: 87 x 13
##                  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, and 7 more variables: birth_year <dbl>,
## #   gender <chr>, homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

Luke Skywalker

eye_color = blue

hair_color = blond

skin_color = fair

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:

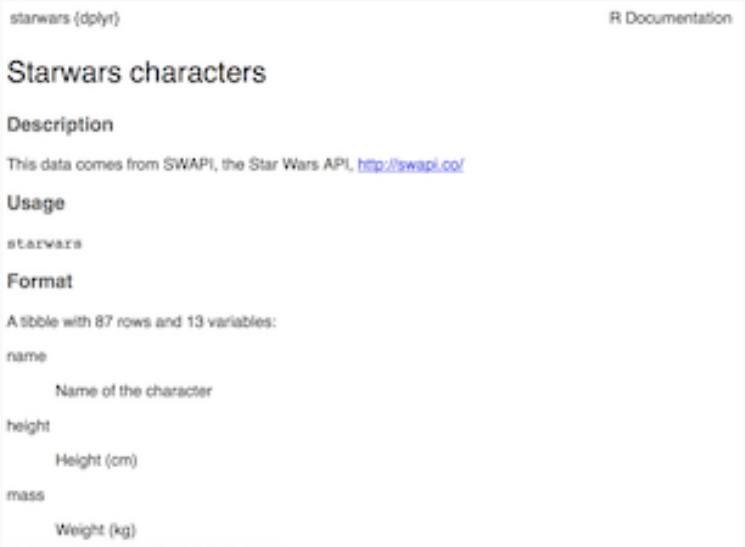
```
glimpse(starwars)
```

```
## 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)

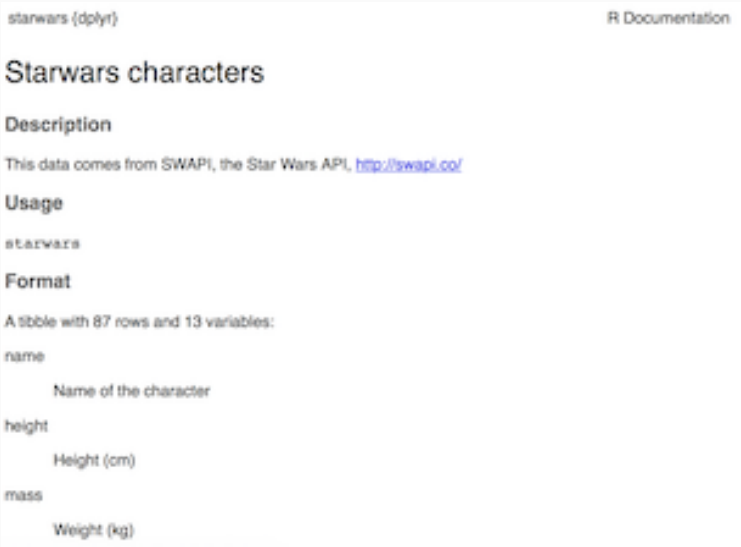
mass

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)

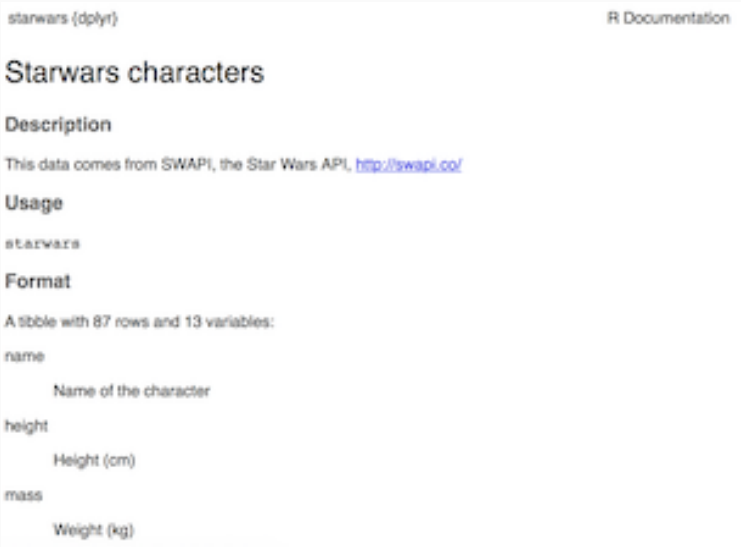
How many rows and columns does this dataset have?

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?

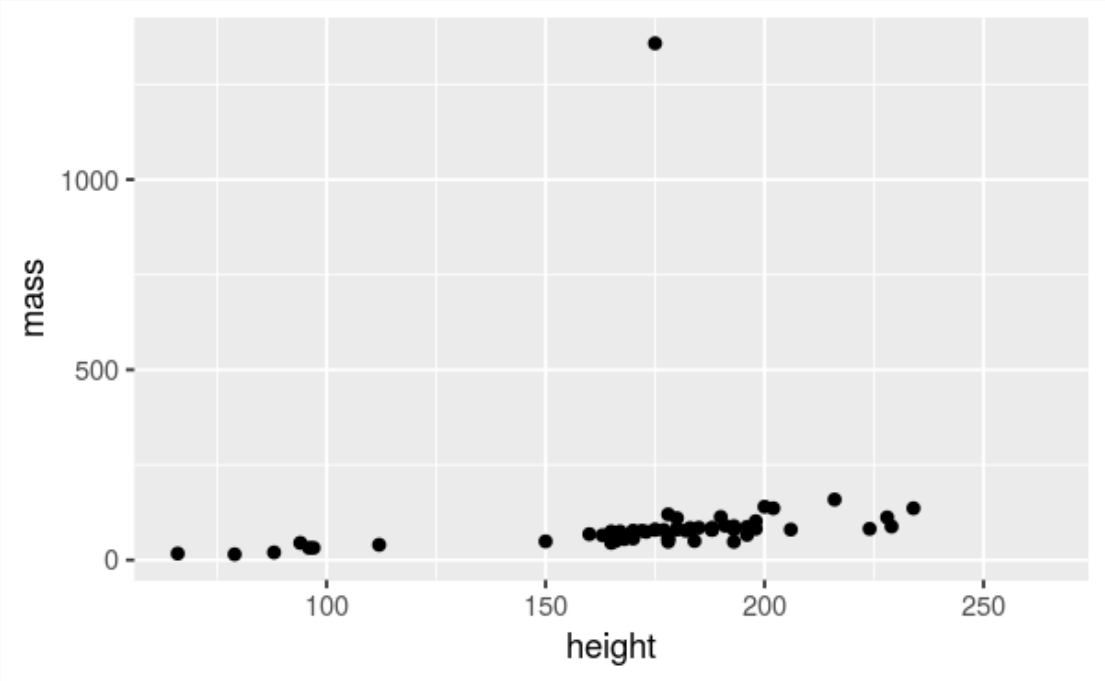
Make a prediction: What relationship do you expect to see between height and mass?

Scatterplots

Mass vs. height (`geom_point()`)

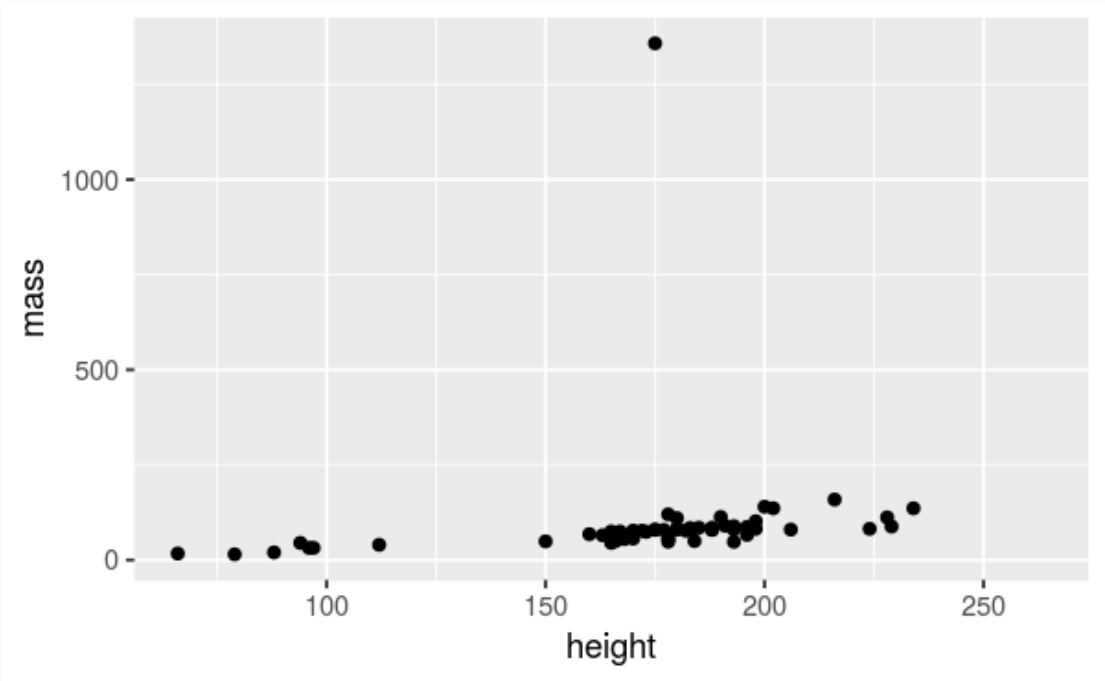
Not all characters have height and mass information (hence 28 of them not plotted)

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



Mass vs. height

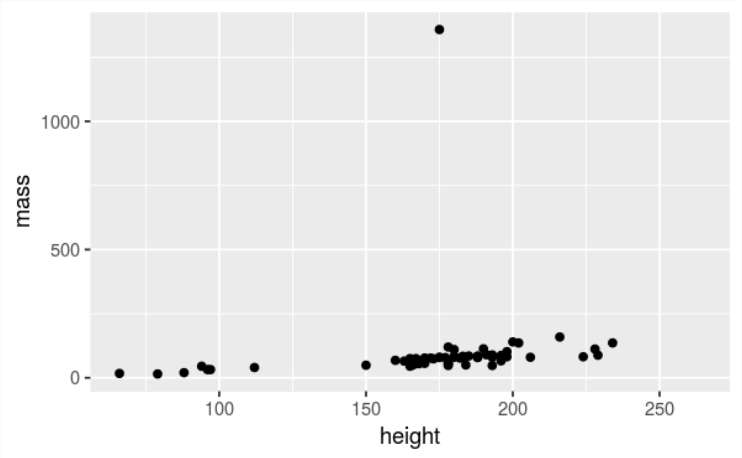
How would you describe this relationship? What other variables would help us understand data points that don't follow the overall trend?



Mass vs. height

Who is the not so tall but really massive character?

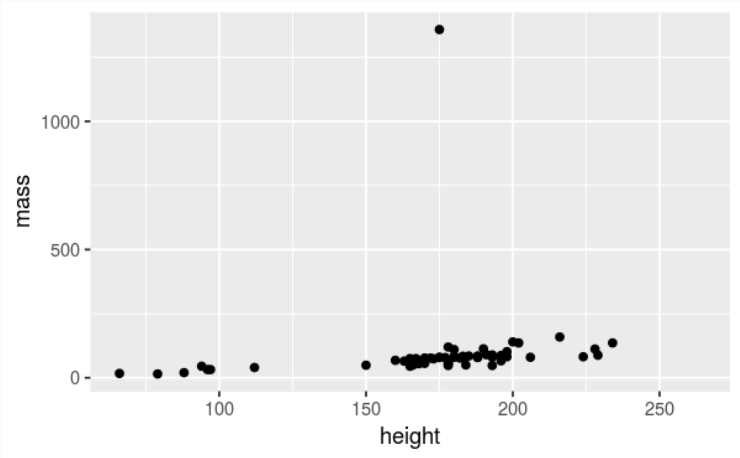
```
ggplot(data = starwars) +
  geom_point(mapping = aes(x = height, y = mass))
```



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

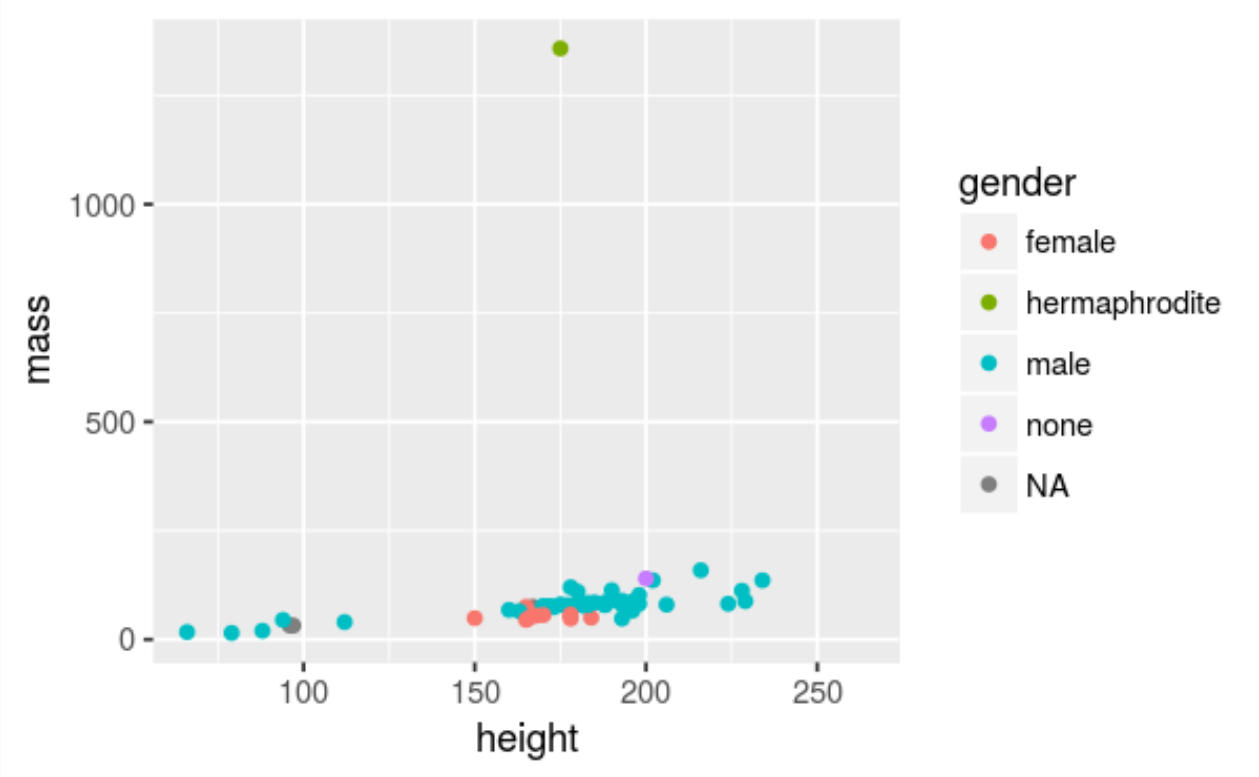
Aesthetics options

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

Faceting options

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

Mass vs. height by gender

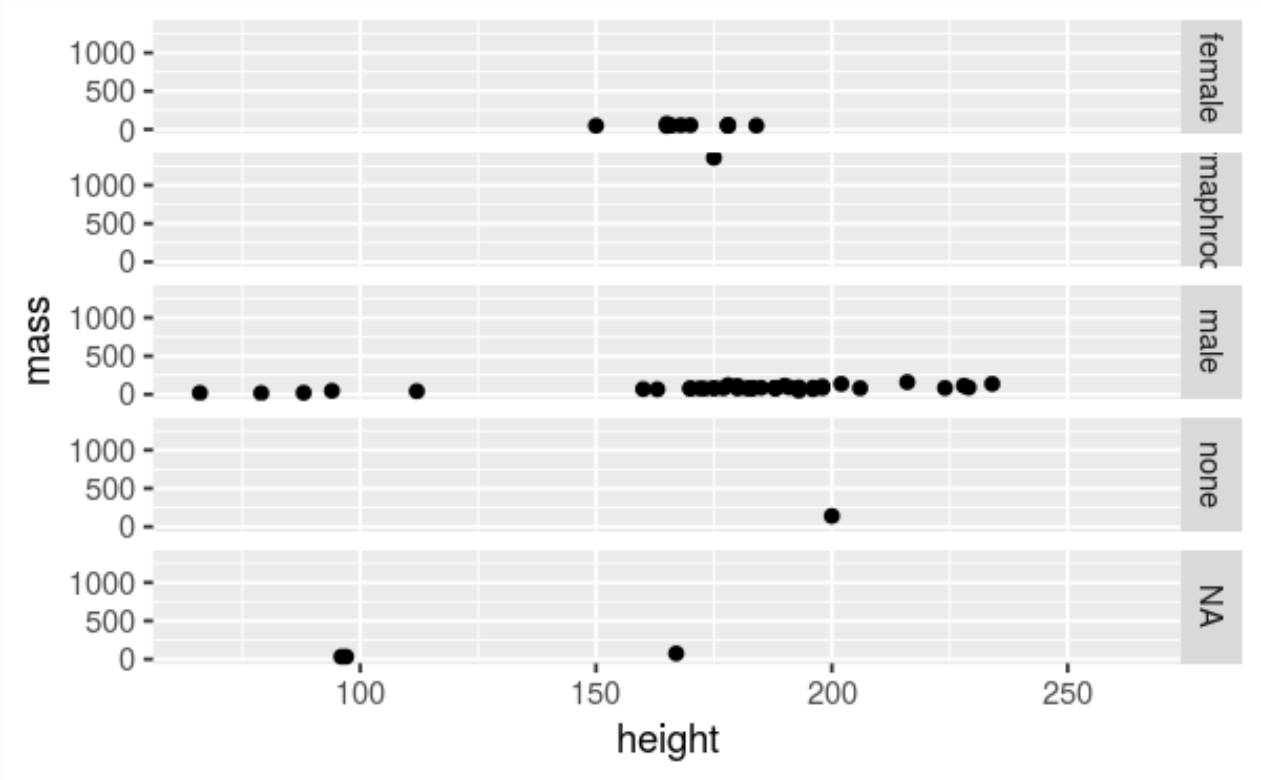
```
ggplot(data = starwars) +
  geom_point(mapping = aes(x = height, y = mass)) +
  facet_grid(. ~ gender)
```

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 ~ .)
```

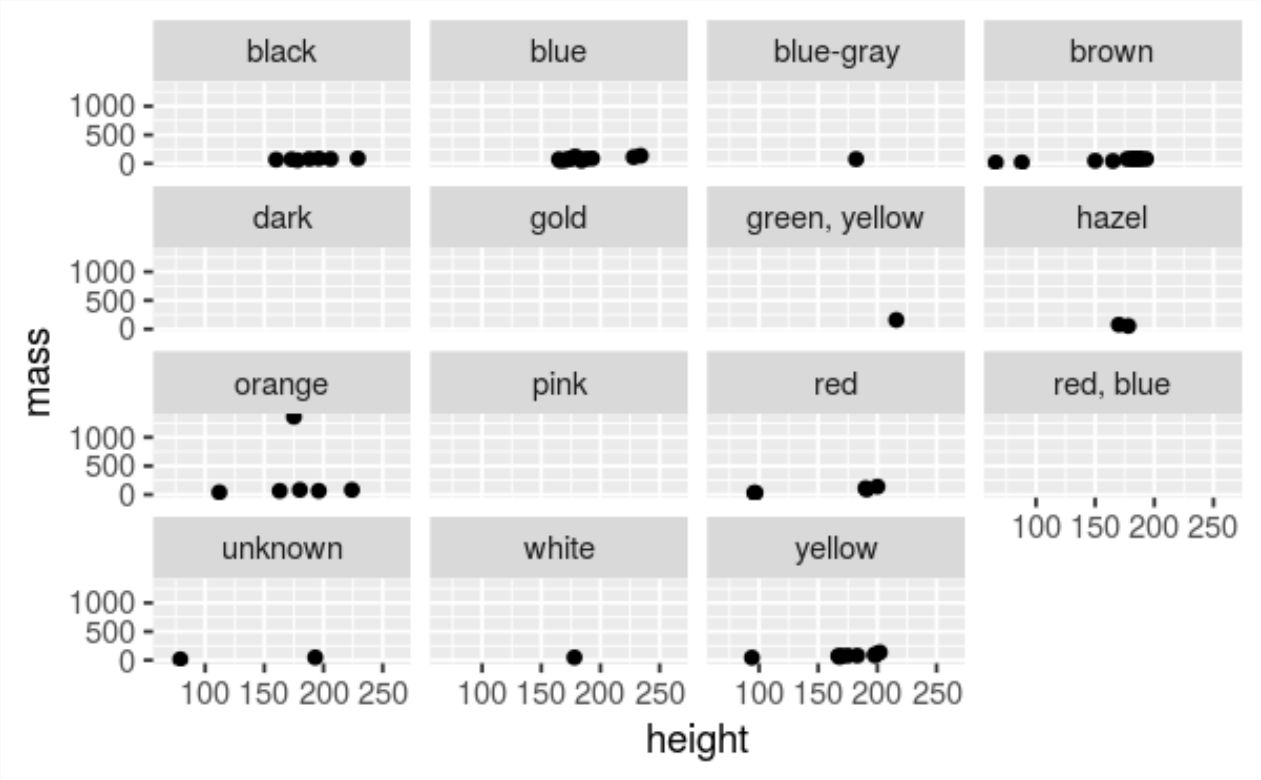


Many ways to facet

```
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

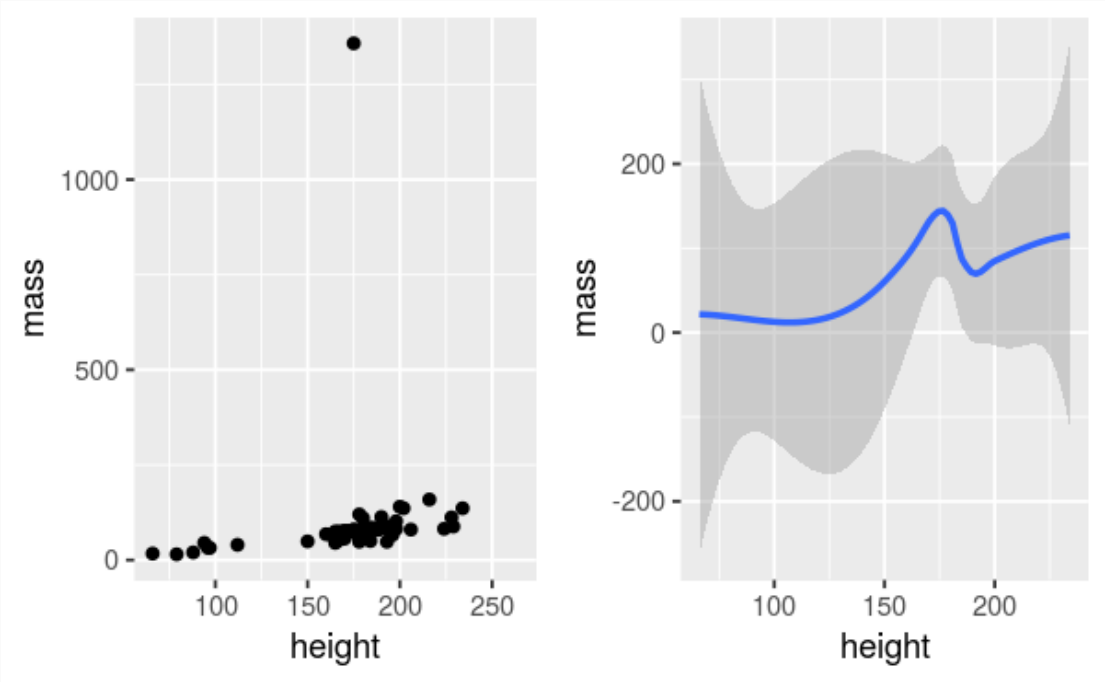
Other geoms

Height vs. mass, take 2

How are these plots similar? How are they different?

Height vs. mass, take 2

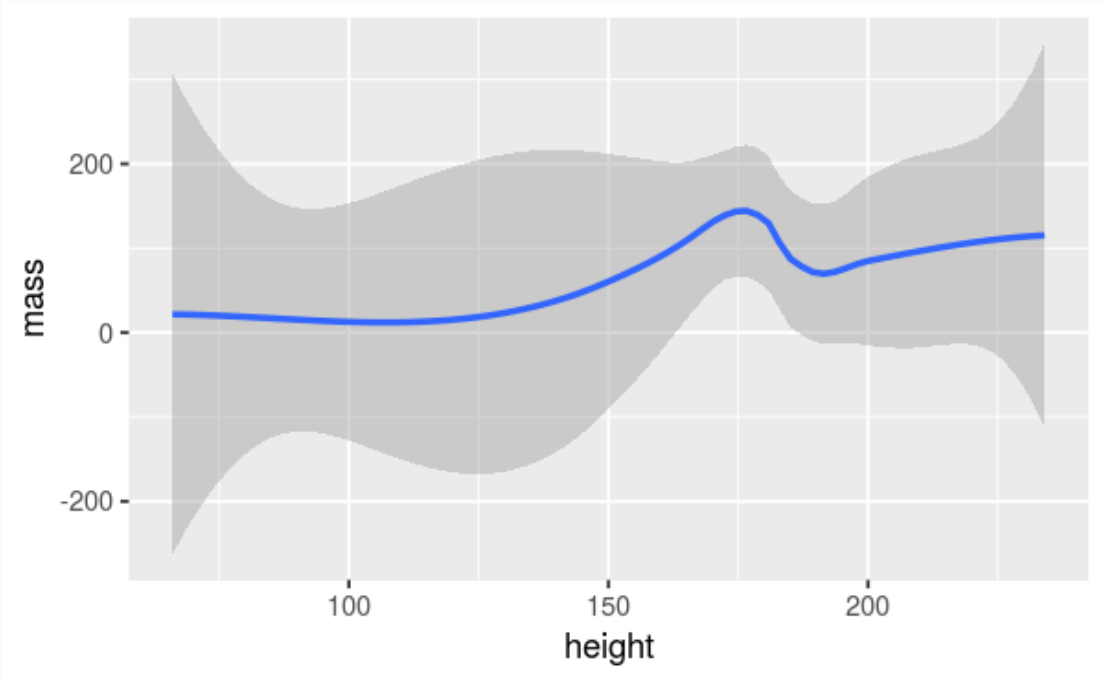
How are these plots similar? How are they different?



geom_smooth

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

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



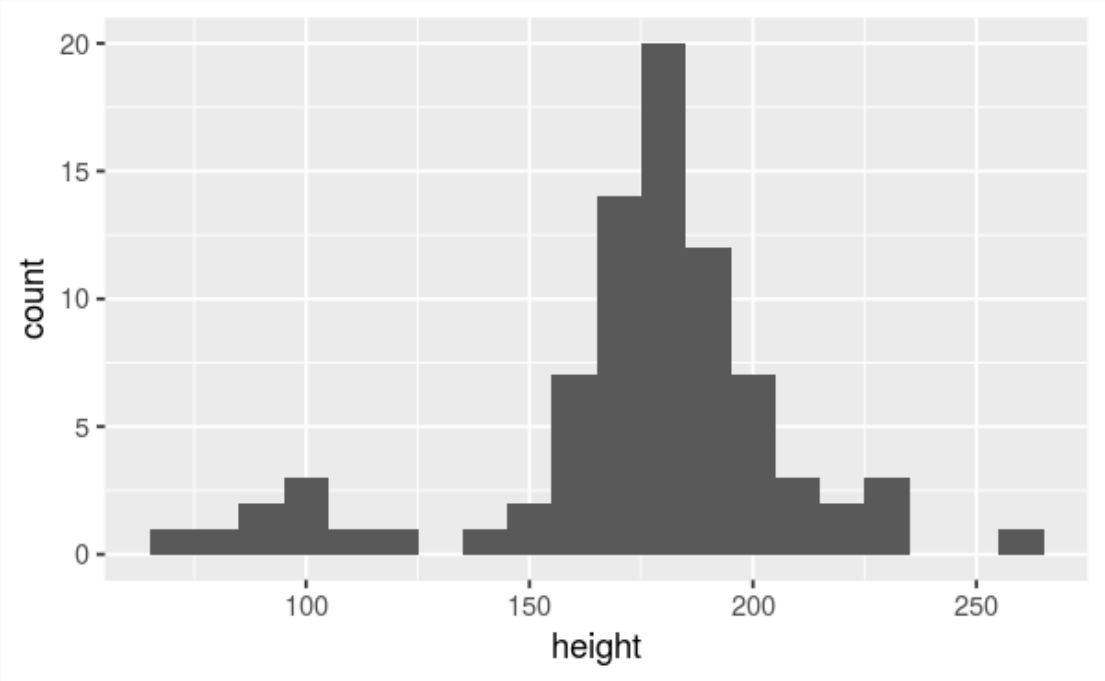
Describing shapes of numerical

- shape:
    - skewness: right-skewed, left-skewed, symmetric (skew is to the side of the longer tail)
    - modality: unimodal, bimodal, multimodal, uniform
- center: mean (`mean`), median (`median`), mode (not always useful)
- spead: range (`range`), standard deviation (`sd`), inter-quartile range (`IQR`)
- unusual observations

Histograms

For numerical variables

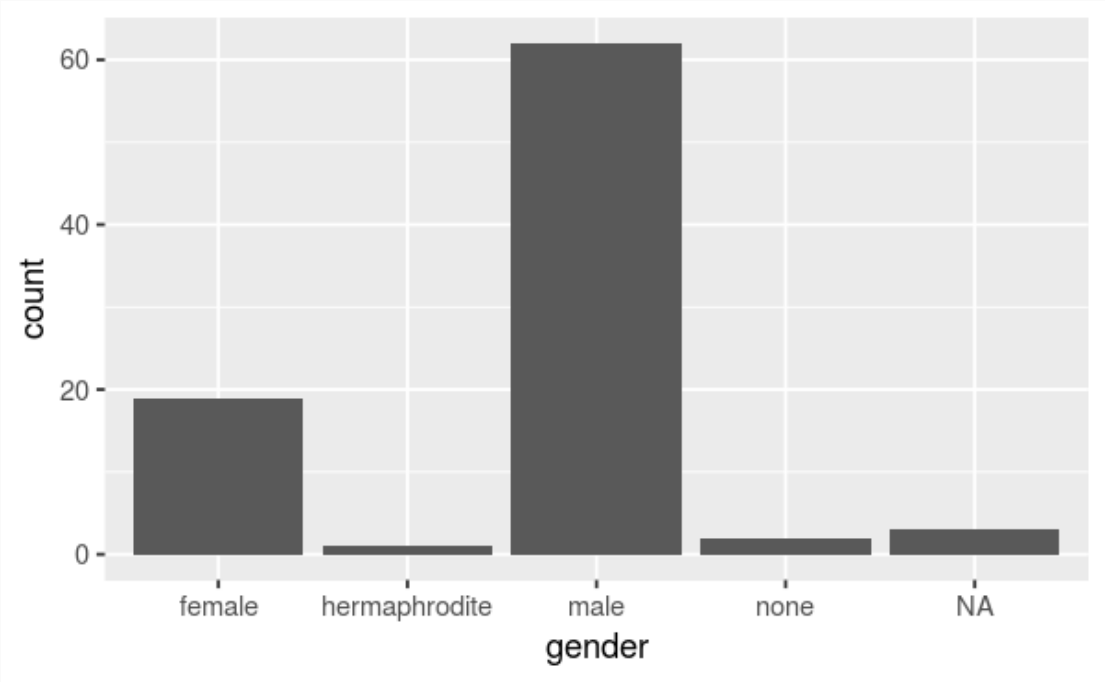
```
ggplot(starwars) +
  geom_histogram(mapping = aes(x = height), binwidth = 10)
```



Bar plots

For categorical variables

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



Group Exercises

Form groups with the neighboring students and complete as many of the following exercises in *R for Data Science* as you can before the class period ends:

- Chapter 3.2.4: exercises 4, 5
- Chapter 3.3.1: exercise 3
- Chapter 3.5.1: exercises 1, 2, 6
- Chapter 3.6.1: exercise 5
- Chapter 3.7.1: exercises 2, 5

At the end of the class period, send me the group `.Rmd` file using Slack.

Credits

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