Section A01
CDS 101 *Introduction to Computational and Data Sciences*

Class 5: Introduction to data and visualization II

May 25, 2018

General

Annoucements

- Course website updated: http://summer18.cds101.com
- Reading 4 from R for Data Science, questions due on May 28th by 5:00pm
  - From chapter 3: section 3.7 through to the end of section 3.10
- Reading 5 from R for Data Science, questions due on May 29th by 9:00am
  - All of chapter 4 (short)
  - All of chapter 5
- Visualization mini-assignment posted, due May 28th @ 11:59pm
- If you cannot access RStudio Server (https://rstudio.cos.gmu.edu) over the weekend, use RStudio Cloud (https://rstudio.cloud) instead

Data visualization as exploration

Basic terms

Variable

A quantity, quality, or property that you can measure.

Value

The state of a variable when you measure it. The value of a variable may change from measurement to measurement.

Observation

A set of measurements made under similar conditions (you usually make all of the measurements in an observation at the same time and on the same object). An observation contains several values, each associated with a different variable.

Basic terms

Tabular data (rectangular data)

A set of values, each associated with a variable and an observation.

Kinds of data

Numerical

Data that is a number, either an *integer* (whole numbers) or a *float* (real numbers). This kind of data is collected from device sensors, through counting and polling, outputs of computational simulations, etc.

Categorical

Groups observations into a set. Categories can be in text form (*strings* or *characters*), for example brand names for a certain kind of product, or numerical, for example labeling city districts by numbers.

Textual

Plain text that is too varied to be treated as a category. Some examples can be full names, the text of a literary work, tweets, etc.

How to describe visualizations

A taxonomy for data graphics

- We can break visualizations down into four basic elements:
  - Visual cues
  - Coordinate system
  - Scale
  - Context

Visual cues

- These are the building blocks of any given visualization.
- Identify 9 separate visual cues.

Cues 1–9

1. **Position** (numerical) where in relation to other things?
2. **Length** (numerical) how big (in one dimension)?
3. **Angle** (numerical) how wide? parallel to something else?
4. **Direction** (numerical) at what slope? In a time series, going up or down?
5. **Shape** (categorical) belonging to which group?
6. **Area** (numerical) how big (in two dimensions)?
7. **Volume** (numerical) how big (in three dimensions)?
8. **Shade** (either) to what extent? how severly?
9. **Color** (either) to what extent? how severly? Beware of red/green color blindness.

Coordinate systems

1. **Cartesian** This is the familiar (x, y)-rectangular coordinate system with two perpendicular axes

2. **Polar**: The radial analog of the Cartesian system with points identified by their radius ρ and angle θ

3. **Geographic**: Locations on the curved surface of the Earth, but represented in a flat two-dimensional plane

Scale

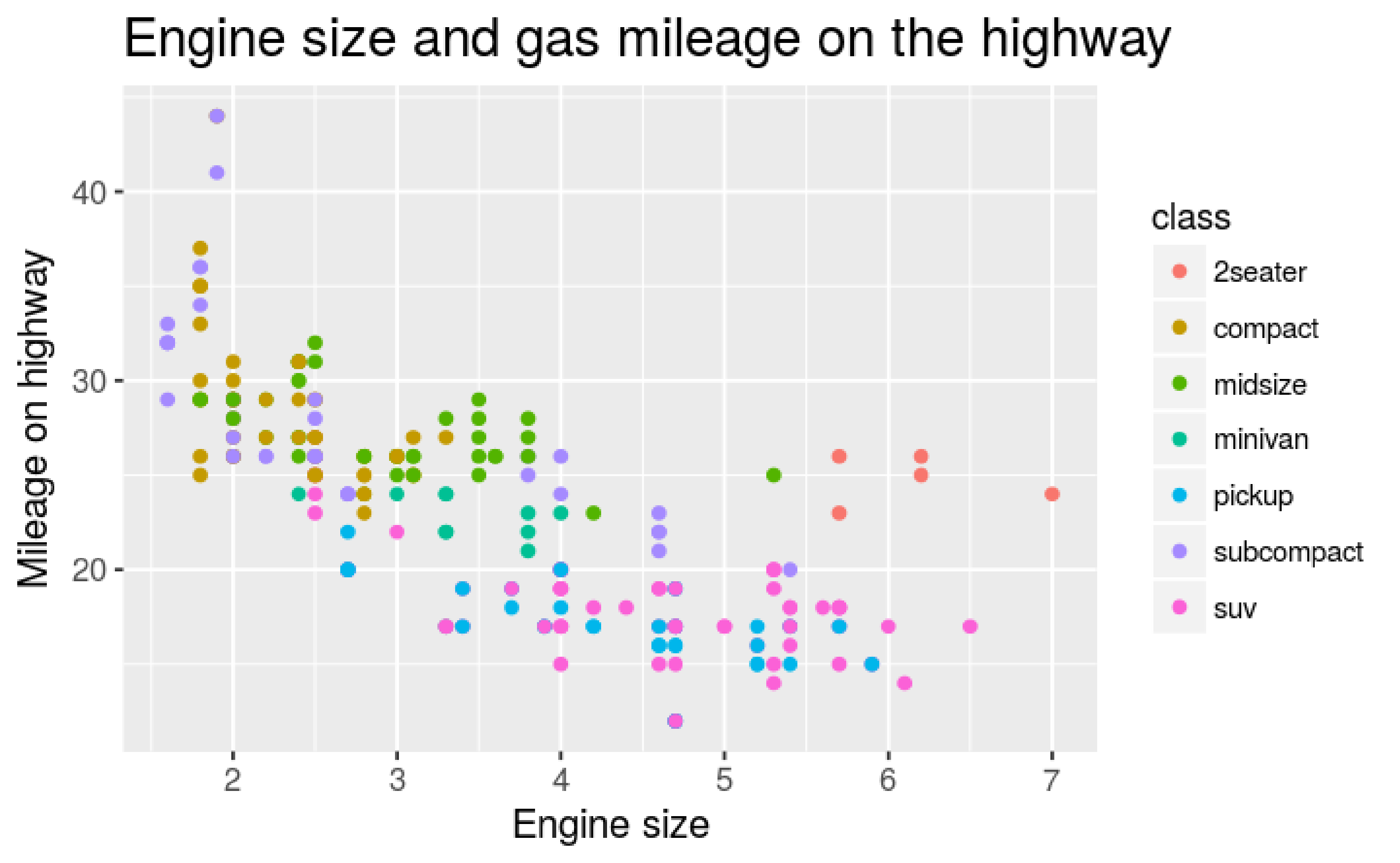
1. **Numeric**: A numeric quantity is most commonly set on a *linear*, *logarithmic*, or *percentage* scale.
2. **Categorical**: A categorical variable may have no ordering or it may be *ordinal* (position in a series).
3. **Time**: A numeric quantity with special properties. Because of the calendar, it can be specified using a series of units (year, month, day). It can also be considered cyclically (years reset back to January, a spring oscillating around a central position).

Context

- Annotations and labels that draw attention to specific parts of a visualization.
  - Titles, subtitles
  - Axes labels that depict scale (tick mark labels) and indiciate the variable
  - Reference points or lines
  - Other markups such as arrows, textboxes, and so on (it's possible to overdo these)

Example plot

How many of the previous elements can you identify in this plot?



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`:

```r
library(tidyverse)
```

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

```r
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
```

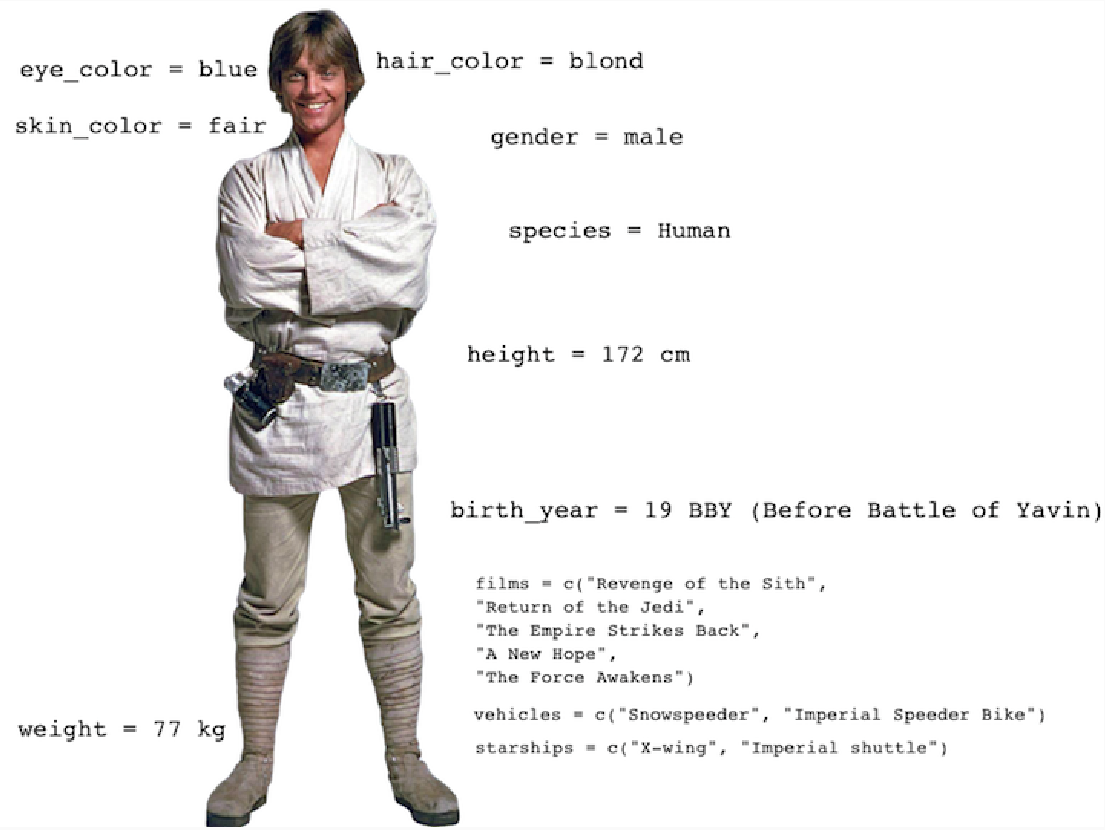
```
## # A tibble: 87 x 13
##    name     height  mass hair_color skin_color eye_color birth_year gender
##    <chr>     <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr>
##  1 Luke Sk…    172    77 blond      fair       blue            19    male
##  2 C-3PO       167    75 <NA>       gold       yellow         112    <NA>
##  3 R2-D2        96    32 <NA>       white, bl… red             33    <NA>
##  4 Darth V…    202   136 none       white      yellow          41.9 male
##  5 Leia Or…    150    49 brown      light      brown           19    female
##  6 Owen La…    178   120 brown, gr… light      blue            52    male
##  7 Beru Wh…    165    75 brown      light      blue            47    female
##  8 R5-D4        97    32 <NA>       white, red red             NA    <NA>
##  9 Biggs D…    183    84 black      light      brown           24    male
## 10 Obi-Wan…    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?

```
## # A tibble: 87 x 13
##    name      height  mass hair_color skin_color eye_color birth_year ge
##    <chr>      <int> <dbl> <chr>      <chr>      <chr>          <dbl> <c
##  1 Luke Sk…    172    77 blond      fair       blue            19   ma
##  2 C-3PO       167    75 <NA>       gold       yellow         112   <N
##  3 R2-D2        96    32 <NA>       white, bl… red             33   <N
##  4 Darth V…    202   136 none       white      yellow          41.9 ma
##  5 Leia Or…    150    49 brown      light      brown           19   fe
##  6 Owen La…    178   120 brown, gr… light      blue            52   ma
##  7 Beru Wh…    165    75 brown      light      blue            47   fe
##  8 R5-D4        97    32 <NA>       white, red red             NA   <N
##  9 Biggs D…    183    84 black      light      brown           24   ma
## 10 Obi-Wan…    182    77 auburn, w… fair       blue-gray       57   ma
## # ... 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:

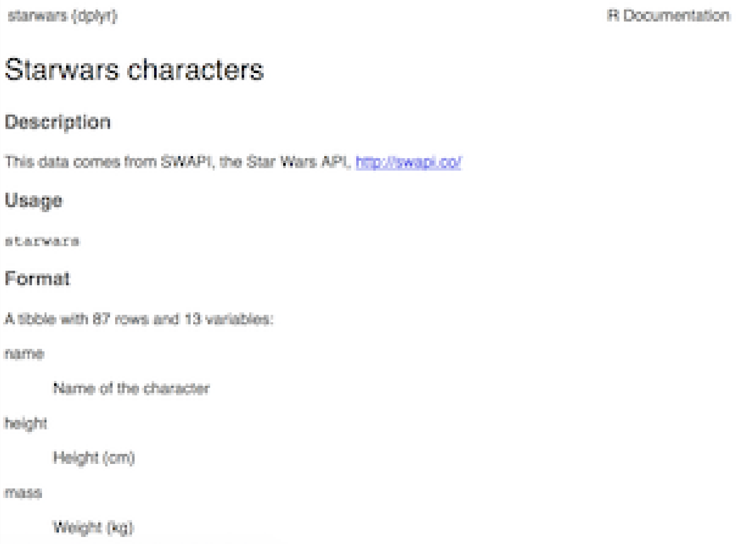
```
glimpse(starwars)
```

```
## 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)

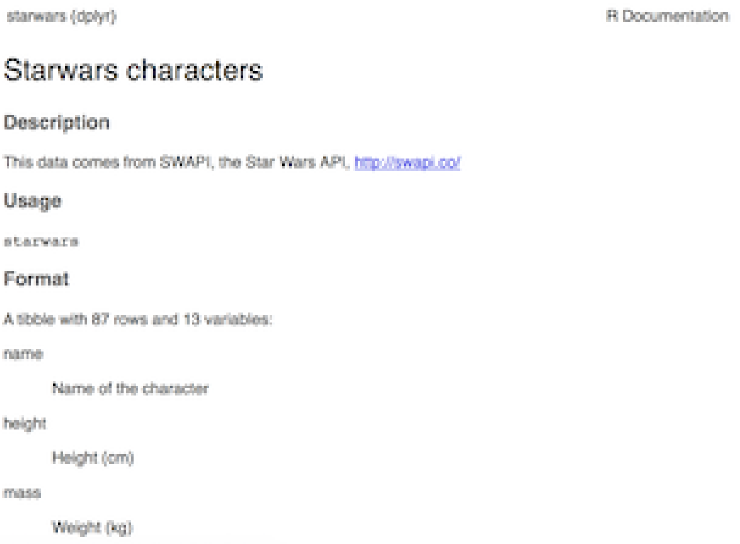
mass

    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)

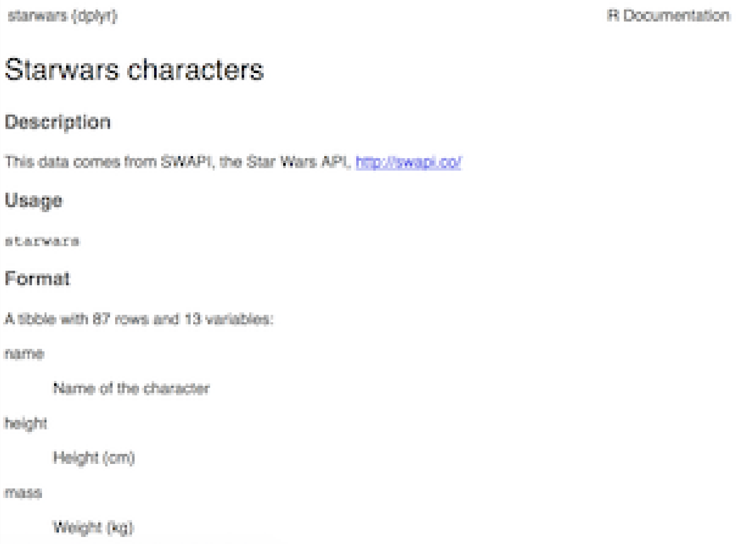
How many rows and columns does this dataset have?

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?

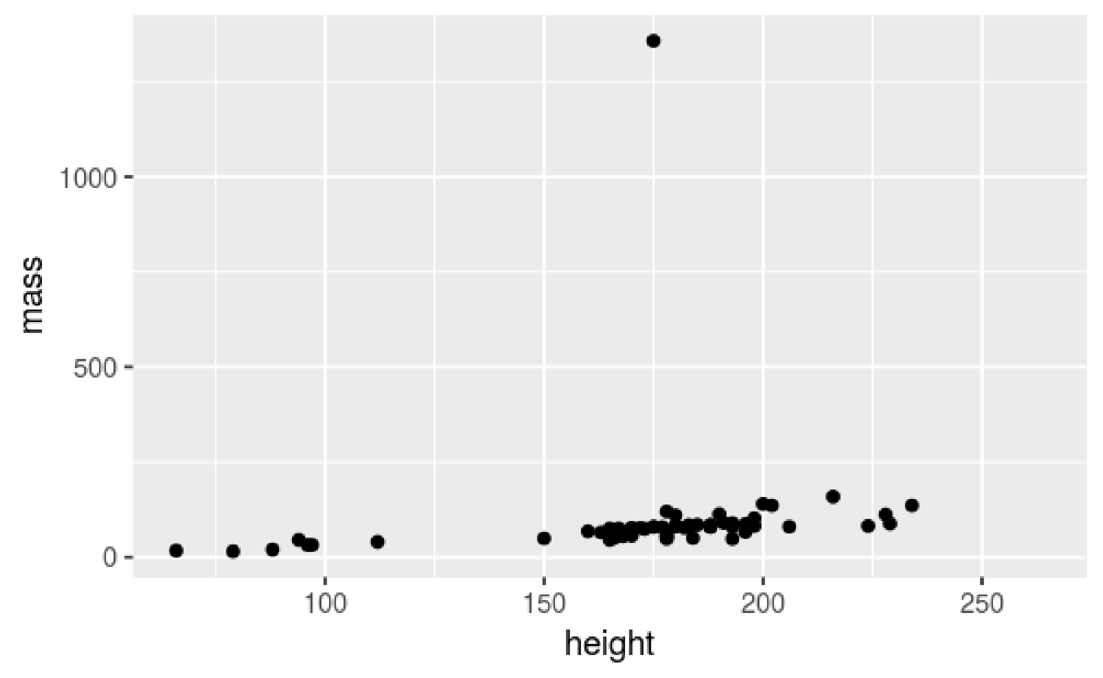
Make a prediction: What relationship do you expect to see between height and mass?

Scatterplots

Mass vs. height (`geom_point()`)

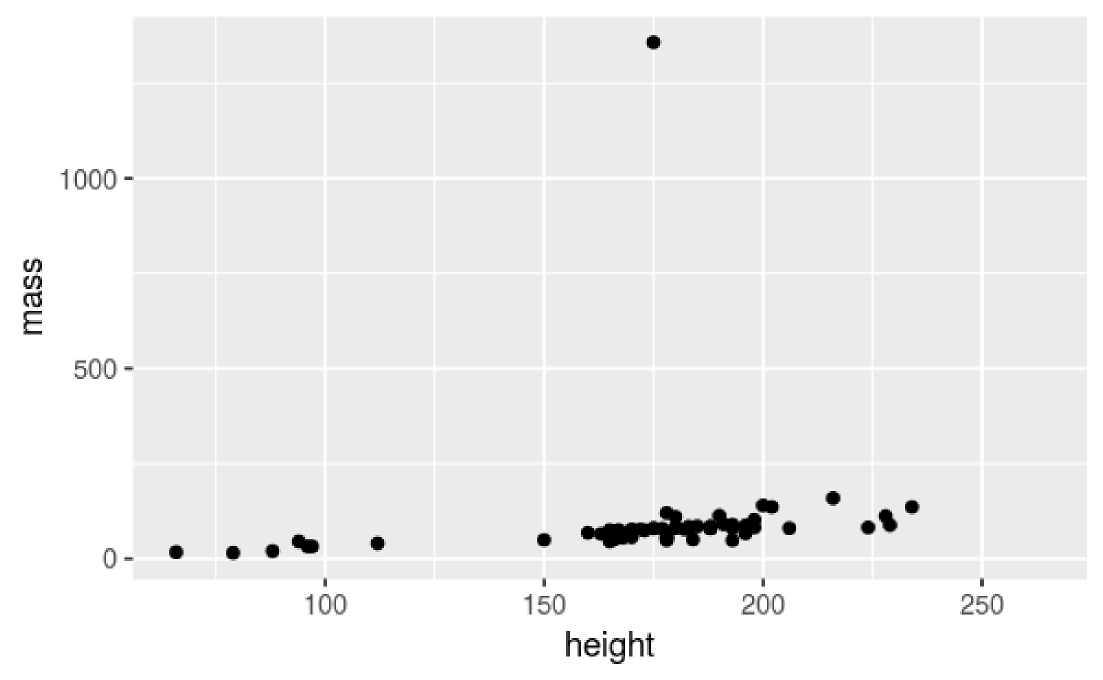
Not all characters have height and mass information (hence 28 of them not plotted)

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



Mass vs. height

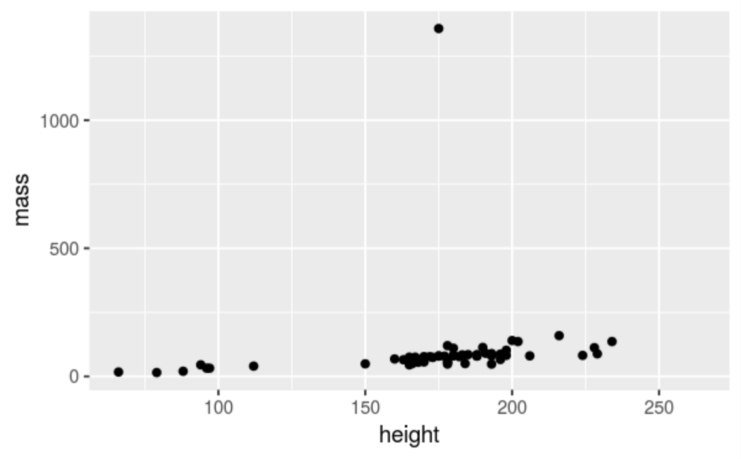
How would you describe this relationship? What other variables would help us understand data points that don't follow the overall trend?



Mass vs. height

Who is the not so tall but really massive character?

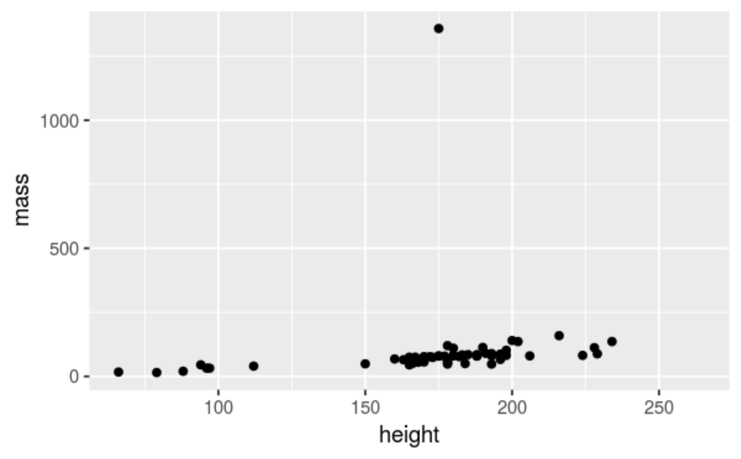
```
ggplot(data = starwars) +
  geom_point(mapping = aes(x = height, y = mass))
```



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

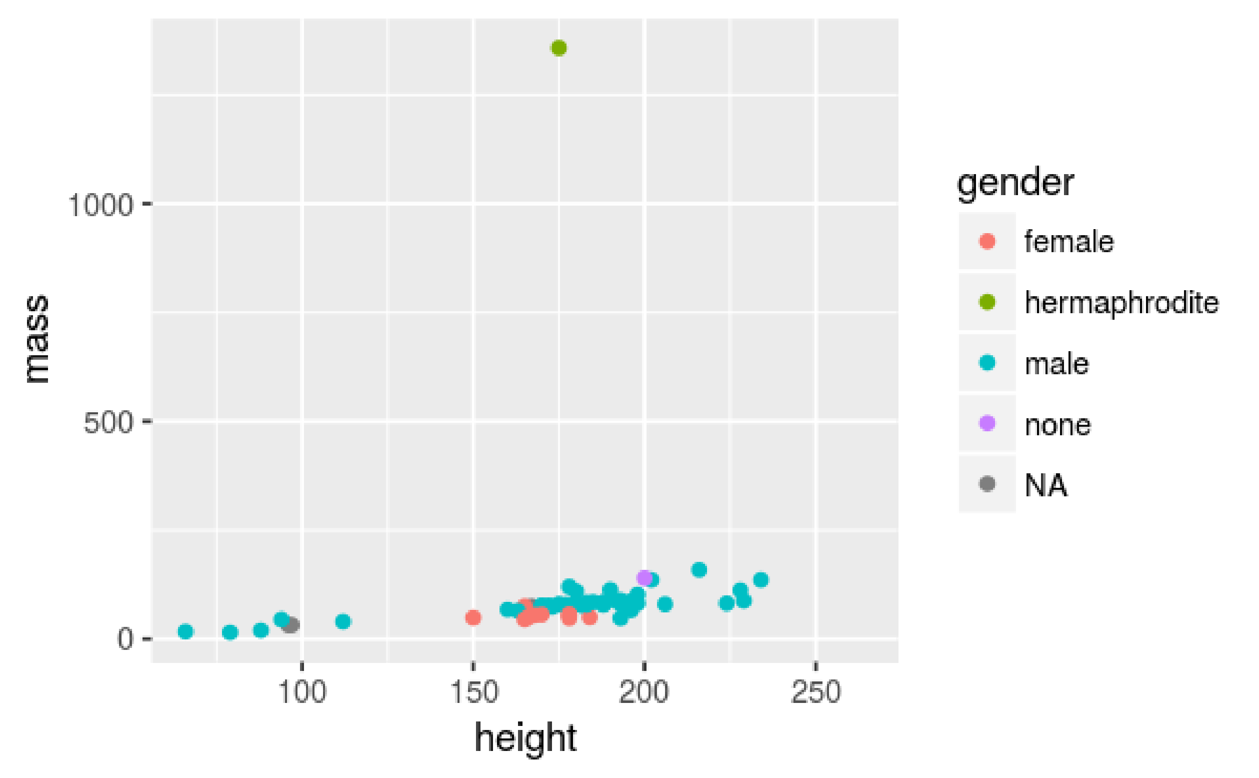
Aesthetics options

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

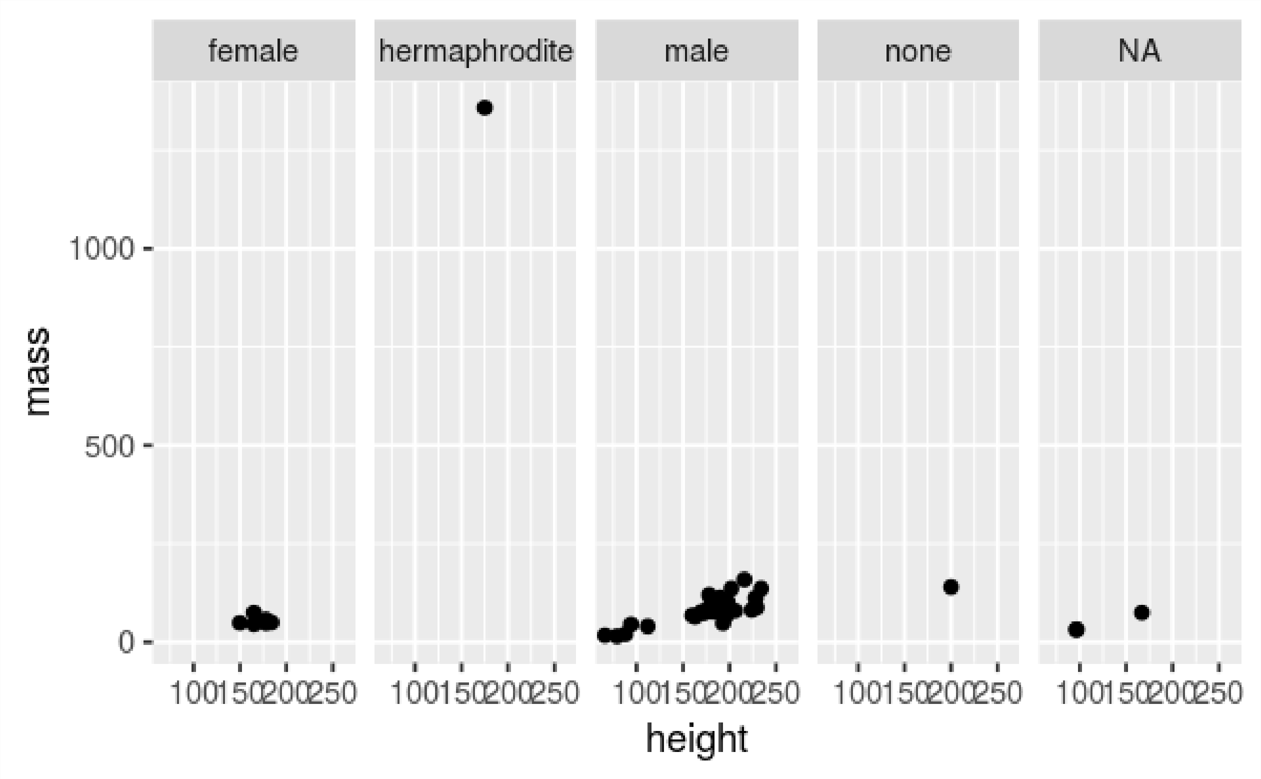
Faceting

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)
```

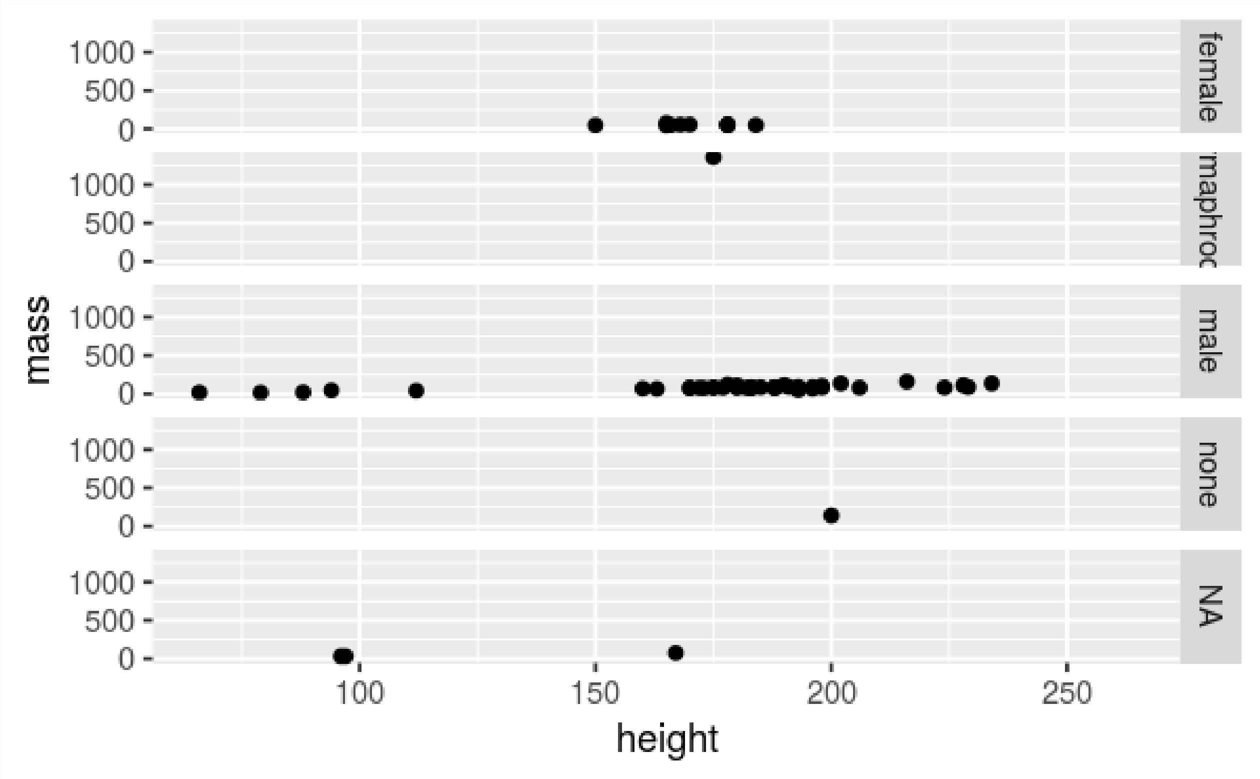


Many ways to facet

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

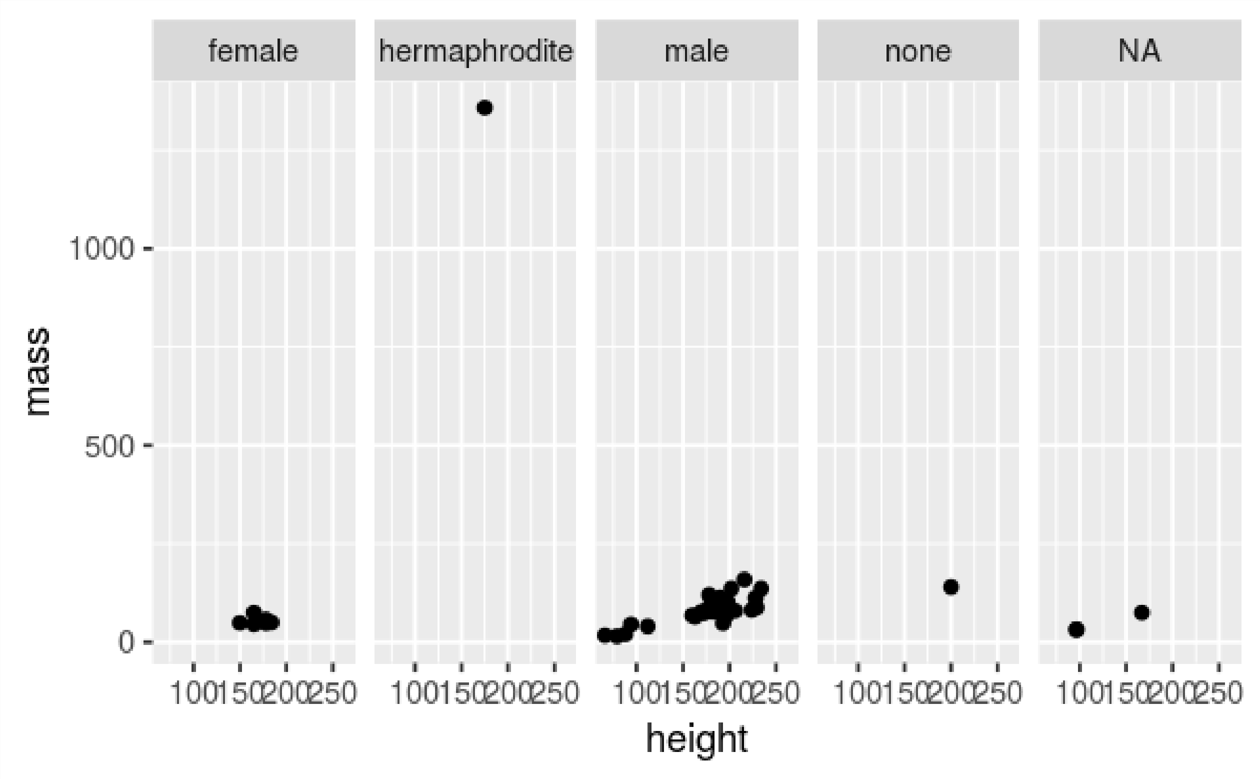
Many ways to facet

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



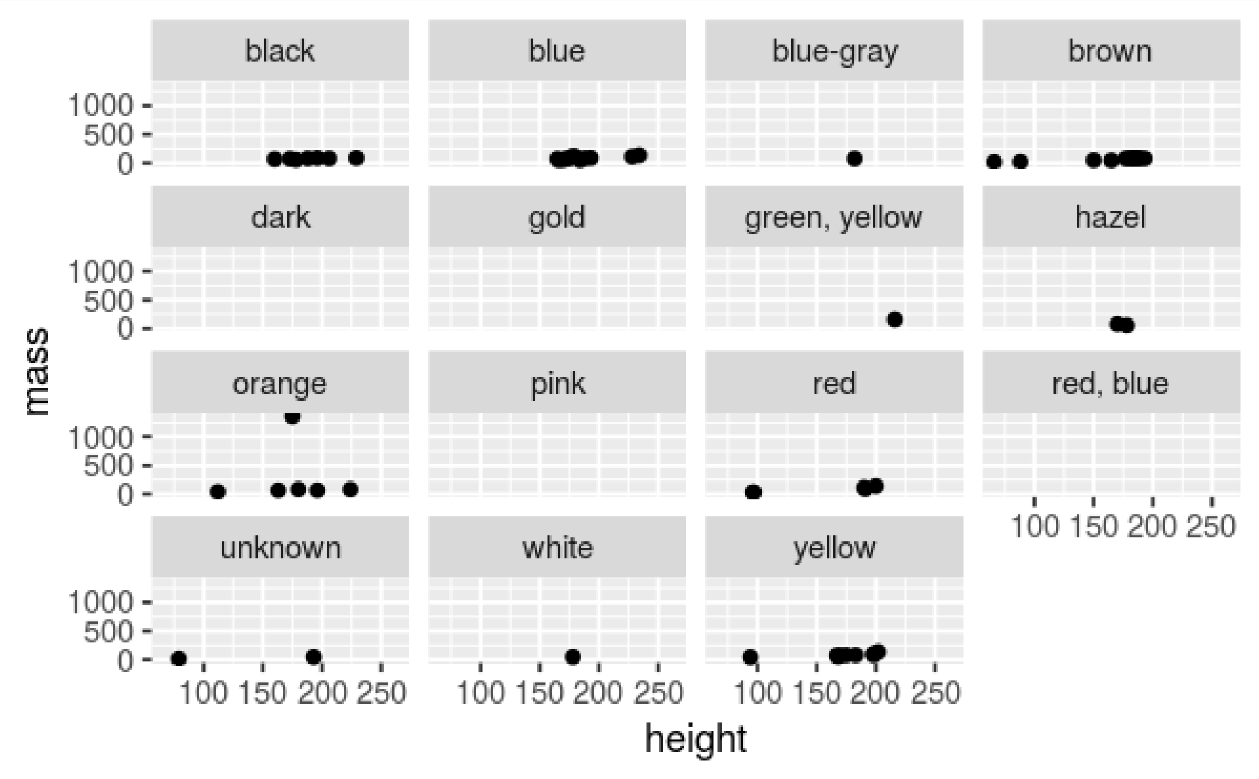
Many ways to facet

```
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

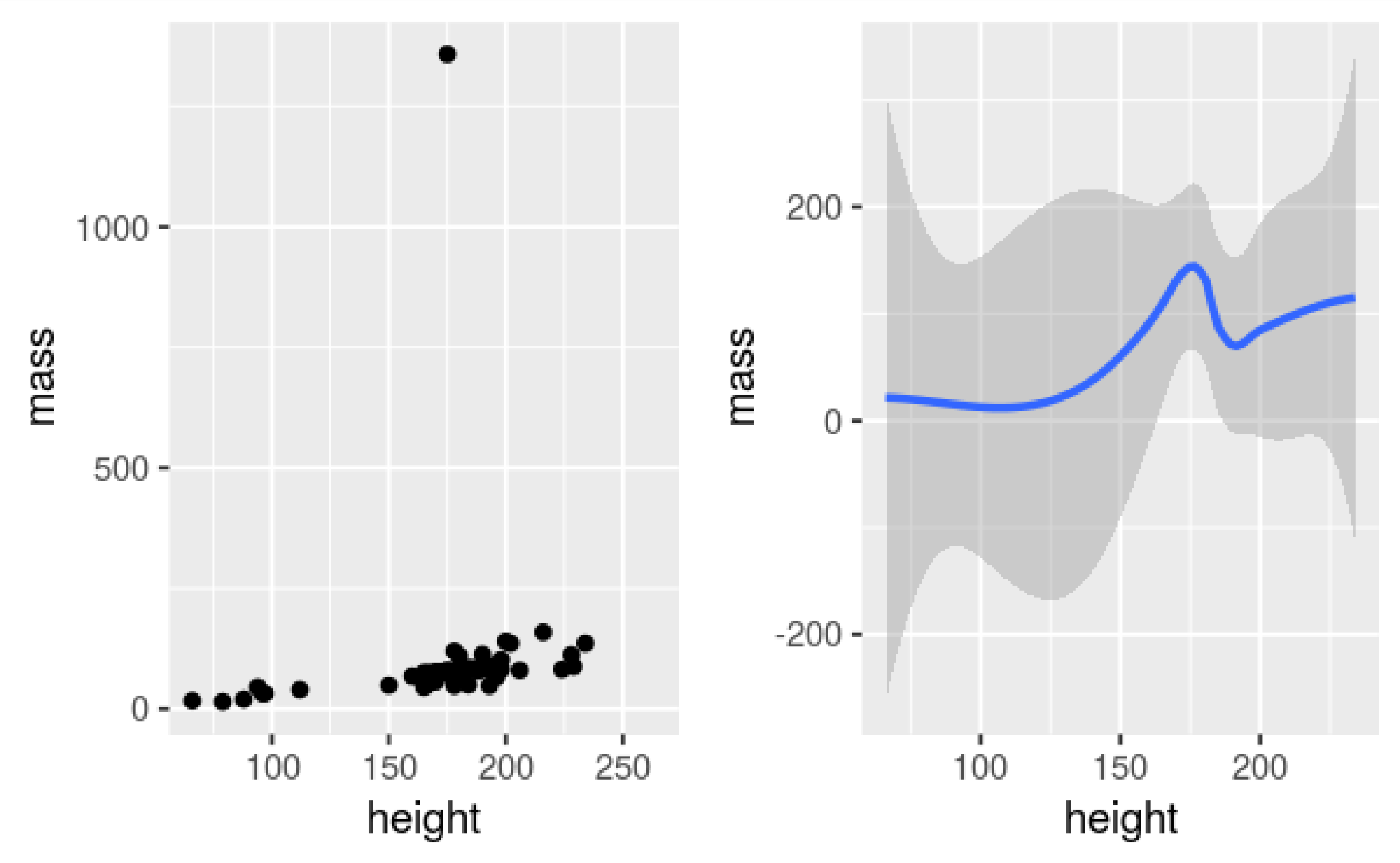
Other geoms

Height vs. mass, take 2

How are these plots similar? How are they different?

Height vs. mass, take 2

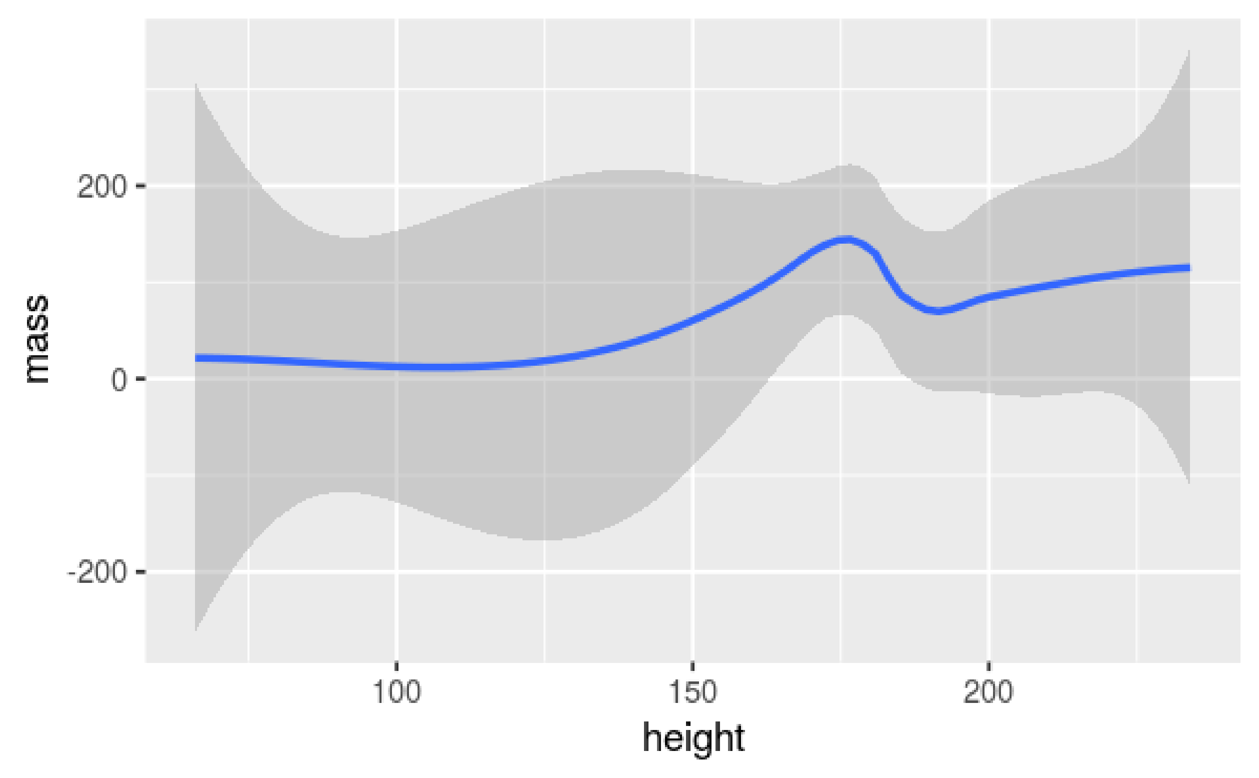
How are these plots similar? How are they different?



geom_smooth

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

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



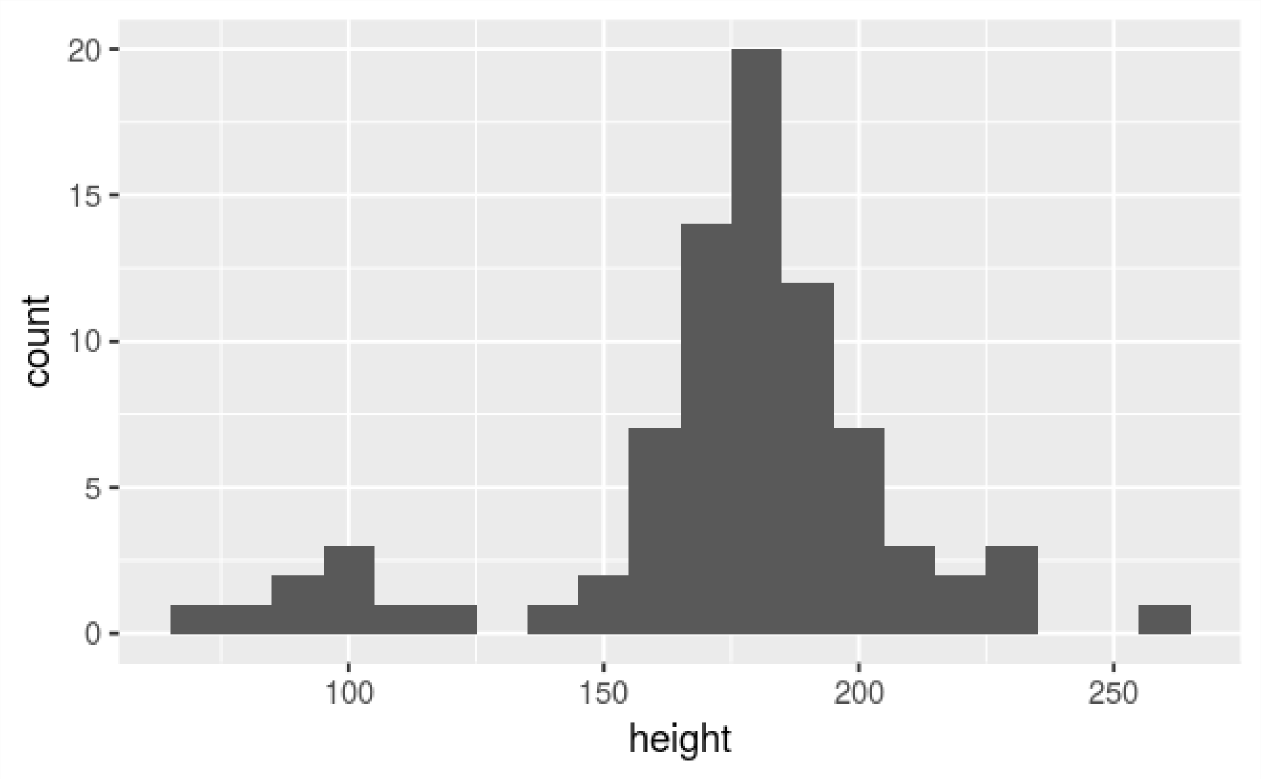
Describing shapes of numerical distributions

- 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)
- spread: range (`range`), standard deviation (`sd`), inter-quartile range (`IQR`)
- unusual observations

Histograms

For numerical variables

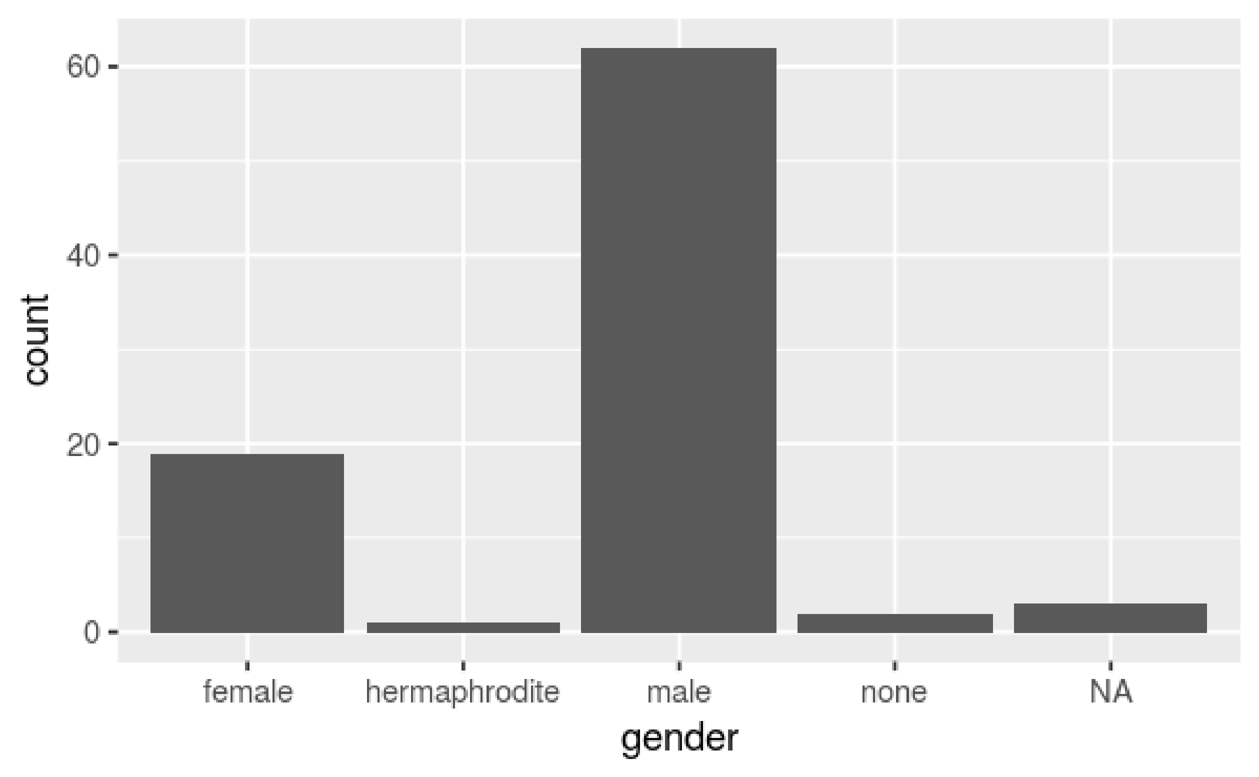
```
ggplot(starwars) +
  geom_histogram(mapping = aes(x = height), binwidth = 10)
```



Bar plots

For categorical variables

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



Credits

These slides were adapted from the following sources:

- Ideas and examples in the section **How to describe visualizations** were adapted from *Modern Data Science with R* by Benjamin Baumer, Daniel Kaplan, and Nicholas Horton, chapter 2.
- Ideas, examples and descriptions from section **Data visualization with ggplot2** onward were adapted from the Fundamentals of data & data visualization slides developed by Mine Çetinkaya-Rundel and made available under the CC BY license.