Data and Visualization

Dr. Maria Tackett

09.03.19

Click for PDF of slides

Announcements

- Regular office hours start this week. Check the course homepage for the office hours schedule
- Lab 01 - due Thursday at 11:59p
- Get to Know You Survey - due TODAY at 11:59p

Check in

- Any questions on material from last time?
- Any questions on the lab?
- Any questions on workflow / course structure?

Exploratory data analysis

What is EDA?

- Exploratory data analysis (EDA) is an approach to analyzing data sets to summarize the main characteristics.
- Often, EDA is visual. That's what we're focusing on today.
- We can also calculate summary statistics and perform data wrangling/manipulation/transformation at (or before) this stage of the analysis. That's what we're focusing in the next class.

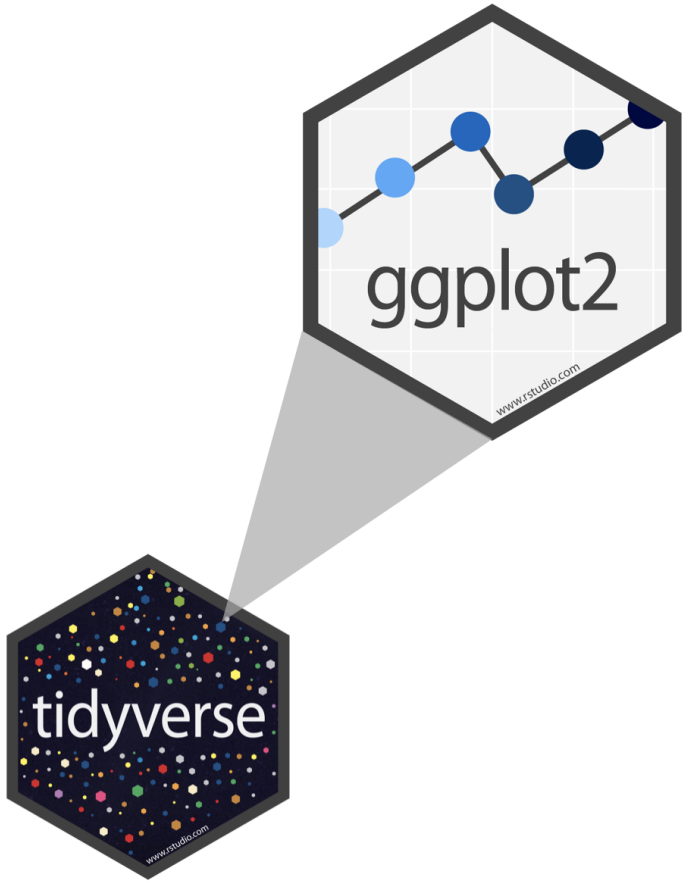
Data visualization

Data visualization

> *"The simple graph has brought more information to the data analyst's mind than any other device." — John Tukey*

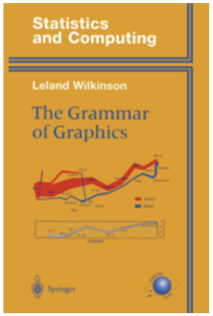
- Data visualization is the creation and study of the visual representation of data.
- There are many tools for visualizing data (R is one of them), and many approaches/systems within R for making data visualizations
  - **ggplot2** is the one we will use

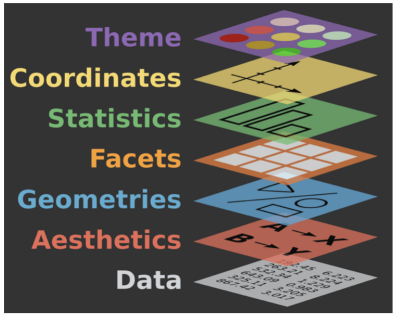
ggplot2 in tidyverse



† Source: BloggoType

- **ggplot2** is tidyverse's data visualization package
- The **gg** in "ggplot2" stands for Grammar of Graphics
- It is inspired by the book **Grammar of Graphics** by Leland Wilkinson †
- A grammar of graphics is a tool that enables us to concisely describe the components of a graphic

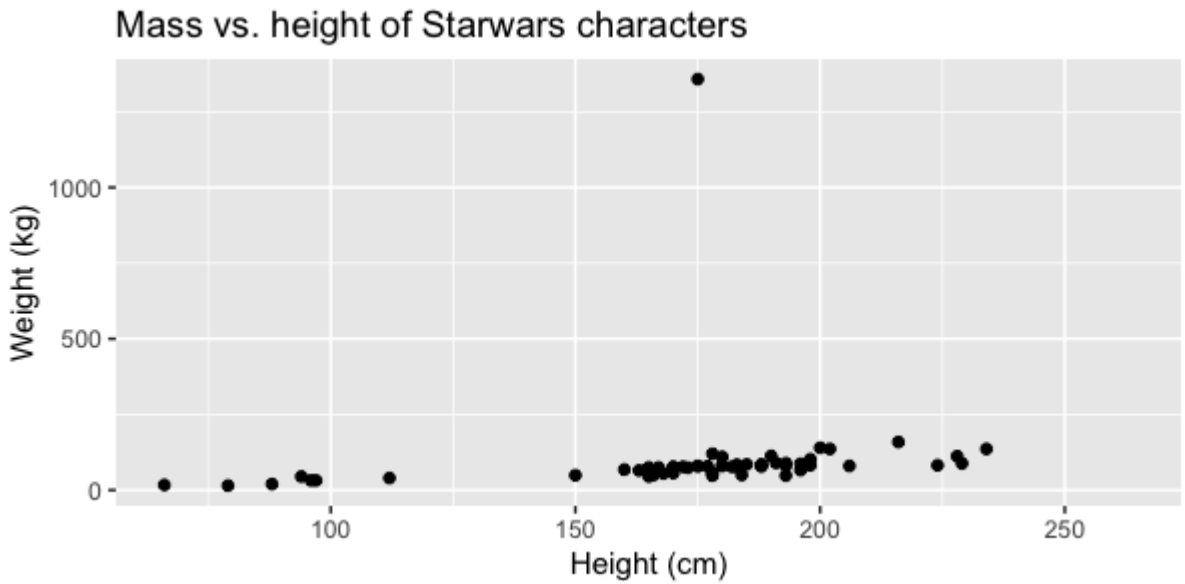




What **functions** are doing the plotting? What is the **dataset** being plotted? Which variable is on the **x-axis**? Which variable is on the **y-axis**? What does the **warning** mean?

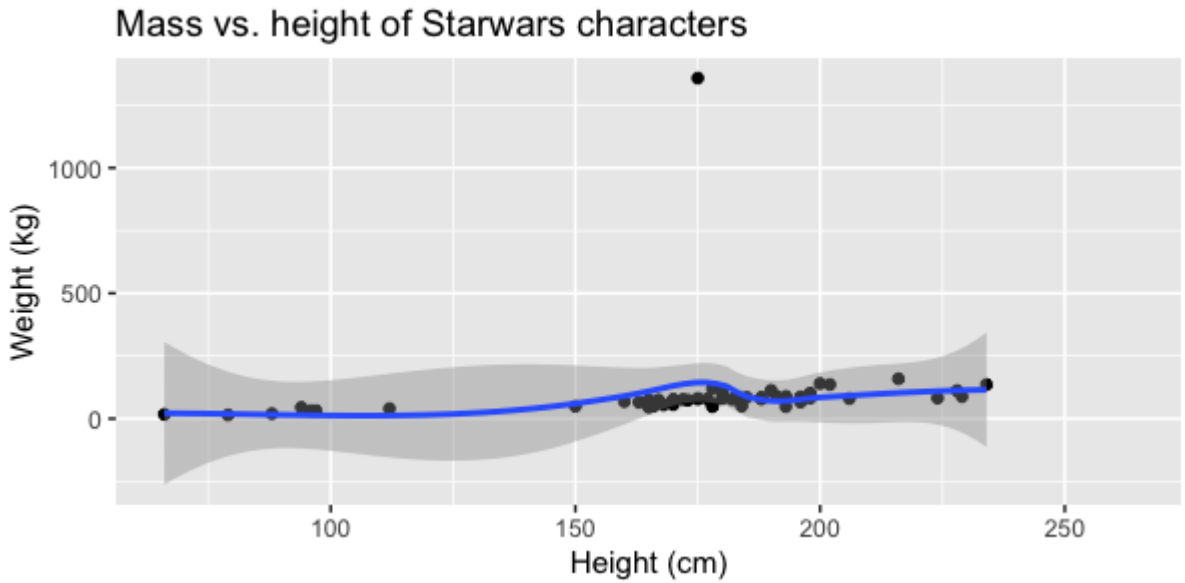
```
ggplot(data = starwars, mapping = aes(x = height, y = mass)) +
  geom_point() +
  labs(title = "Mass vs. height of Starwars characters",
       x = "Height (cm)", y = "Weight (kg)")
```

```
## Warning: Removed 28 rows containing missing values (geom_point).
```



What does **`geom_smooth()`** do? In other words, what changed between the previous plot and this one?

```
ggplot(data = starwars, mapping = aes(x = height, y = mass)) +
  geom_point() +
  geom_smooth() +
  labs(title = "Mass vs. height of Starwars characters",
       x = "Height (cm)", y = "Weight (kg)")
```



Hello ggplot2!

- **`ggplot()`** is the main function in ggplot2 and plots are constructed in layers
- The structure of the code for plots can often be summarized as

```
ggplot +
  geom_xxx
```

or, more precisely

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

- To use ggplot2 functions, first load tidyverse

```
library(tidyverse)
```

- For help with the ggplot2, see ggplot2.tidyverse.org

Visualizing Star Wars

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>
```

Dataset terminology

```
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>
```

- Each row is an observation
- Each column is a variable

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`** of the data:

```
glimpse(starwars)
```

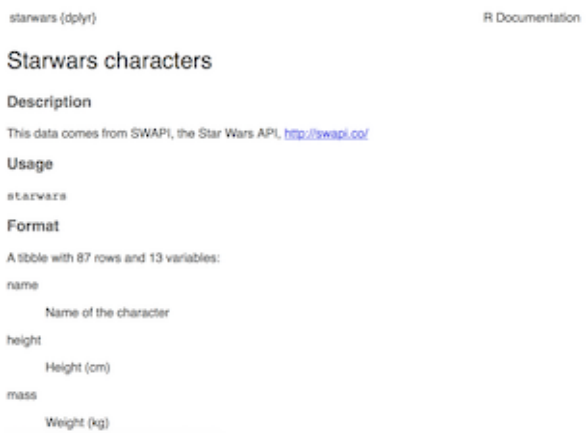
```
## Observations: 87
## Variables: 13
## $ name       <chr> "Luke Skywalker", "C-3PO", "R2-D2", "Darth Vader", "L…
## $ height     <int> 172, 167, 96, 202, 150, 178, 165, 97, 183, 182, 188, …
## $ mass       <dbl> 77.0, 75.0, 32.0, 136.0, 49.0, 120.0, 75.0, 32.0, 84.…
## $ hair_color <chr> "blond", NA, NA, "none", "brown", "brown, grey", "bro…
## $ skin_color <chr> "fair", "gold", "white, blue", "white", "light", "lig…
## $ 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", N…
## $ homeworld  <chr> "Tatooine", "Tatooine", "Naboo", "Tatooine", "Alderaa…
## $ species    <chr> "Human", "Droid", "Droid", "Human", "Human", "Human",…
## $ films      <list> [<"Revenge of the Sith", "Return of the Jedi", "The …
## $ vehicles   <list> [<"Snowspeeder", "Imperial Speeder Bike">, <>, <>, <…
## $ starships  <list> [<"X-wing", "Imperial shuttle">, <>, <>, "TIE Advanc…
```

What's in the Star Wars data?

How many rows and columns does this dataset have? What does each row represent? What does each column represent?

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

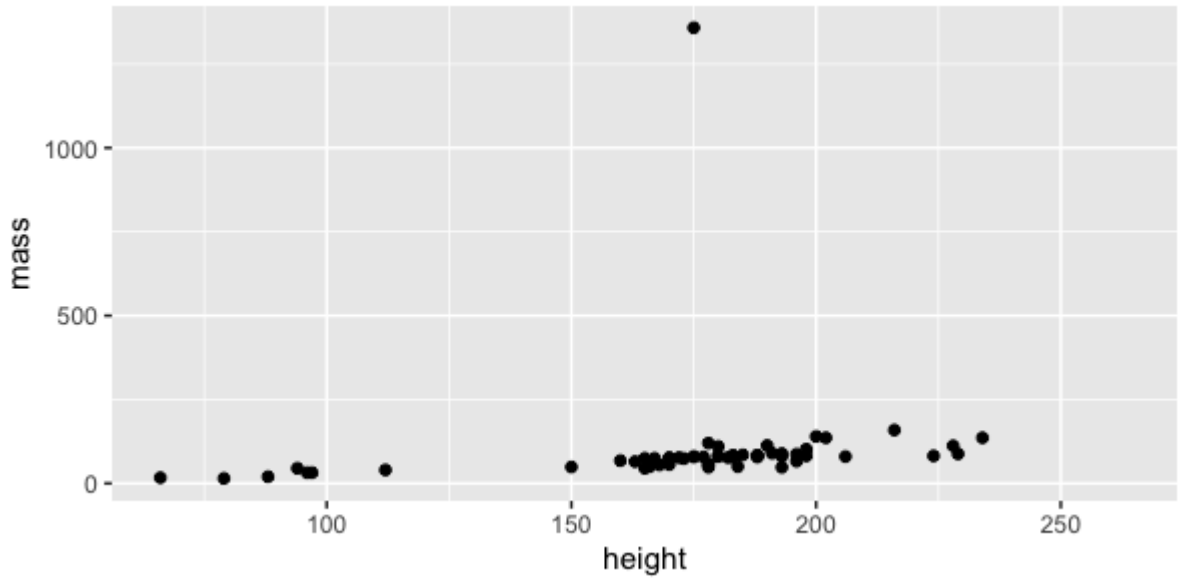
height
: Height (cm)

mass
: Weight (kg)

Mass vs. height

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

```
## Warning: Removed 28 rows containing missing values (geom_point).
```



What's that warning?

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

```
## Warning: Removed 28 rows containing missing values (geom_point).
```

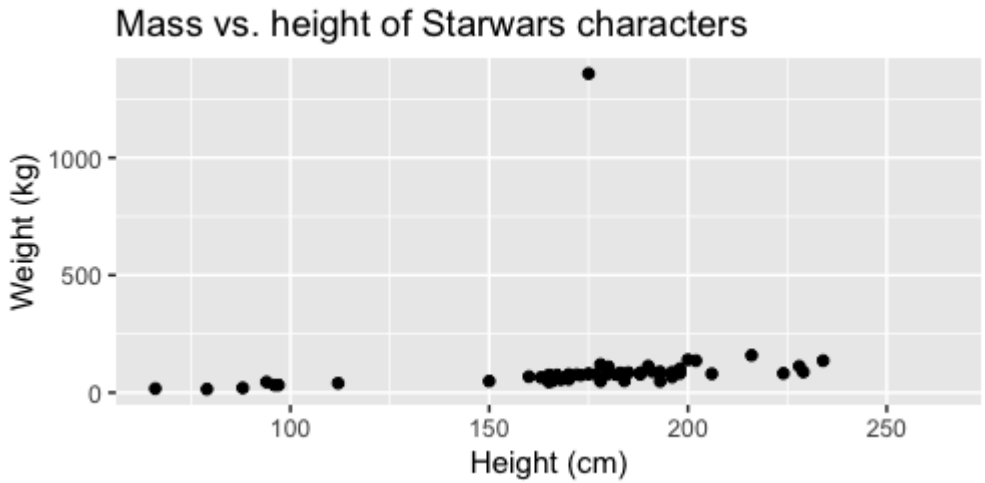
- Going forward I'll suppress the warning to save space on the slides, but it's important to note it
- To suppress warning:

```
{r code-chunk-label, warning=FALSE}
```

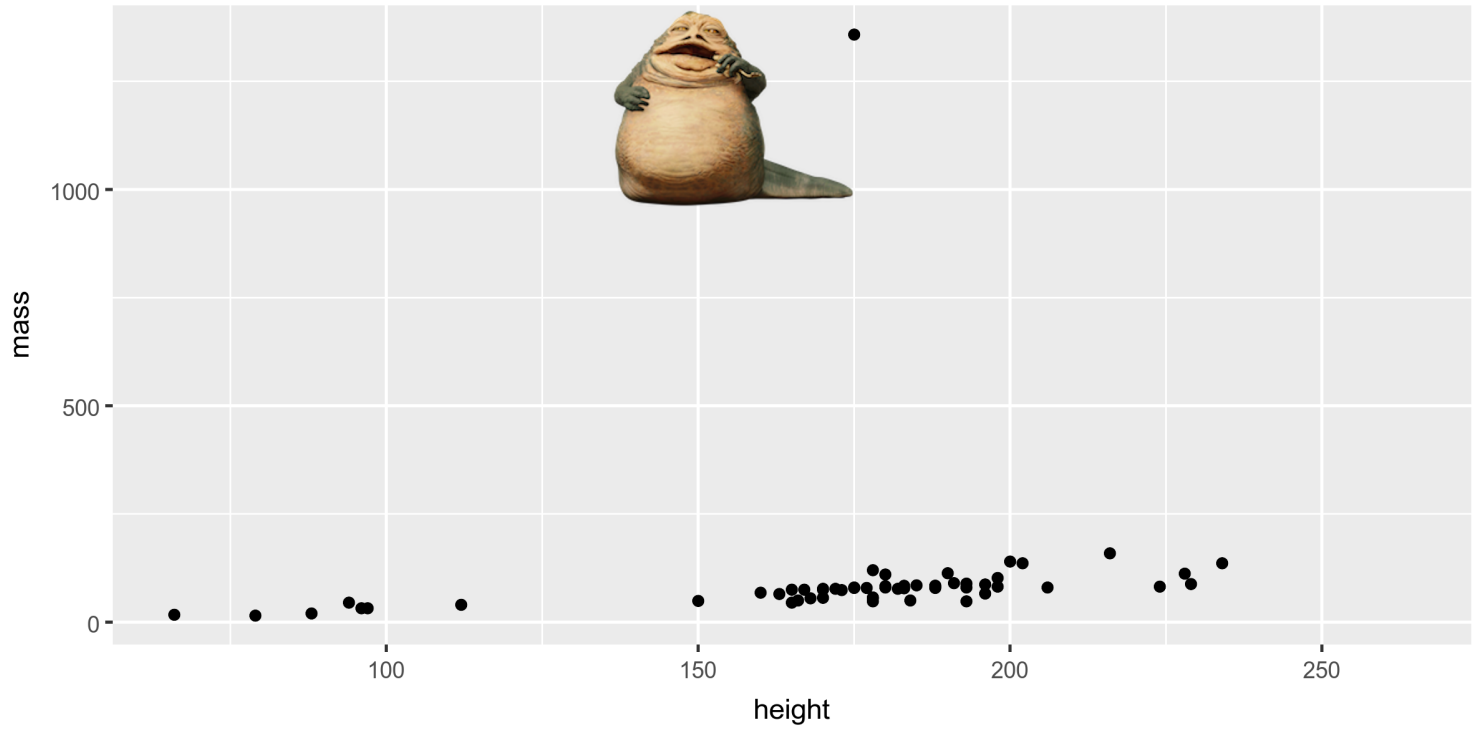
Mass vs. height

How would you describe this **relationship**? What other variables would help us understand data points that don't follow the overall trend? Who is the not so tall but really heavy character?

```
ggplot(data = starwars, mapping = aes(x = height, y = mass)) +
  geom_point() +
  labs(title = "Mass vs. height of Starwars characters",
       x = "Height (cm)", y = "Weight (kg)")
```



Jabba!



Additional variables

We can map additional variables to various features of the plot:

- **aesthetics**
  - shape
  - color
  - size
  - alpha (transparency)
- **faceting**: small multiples displaying different subsets

Aesthetics

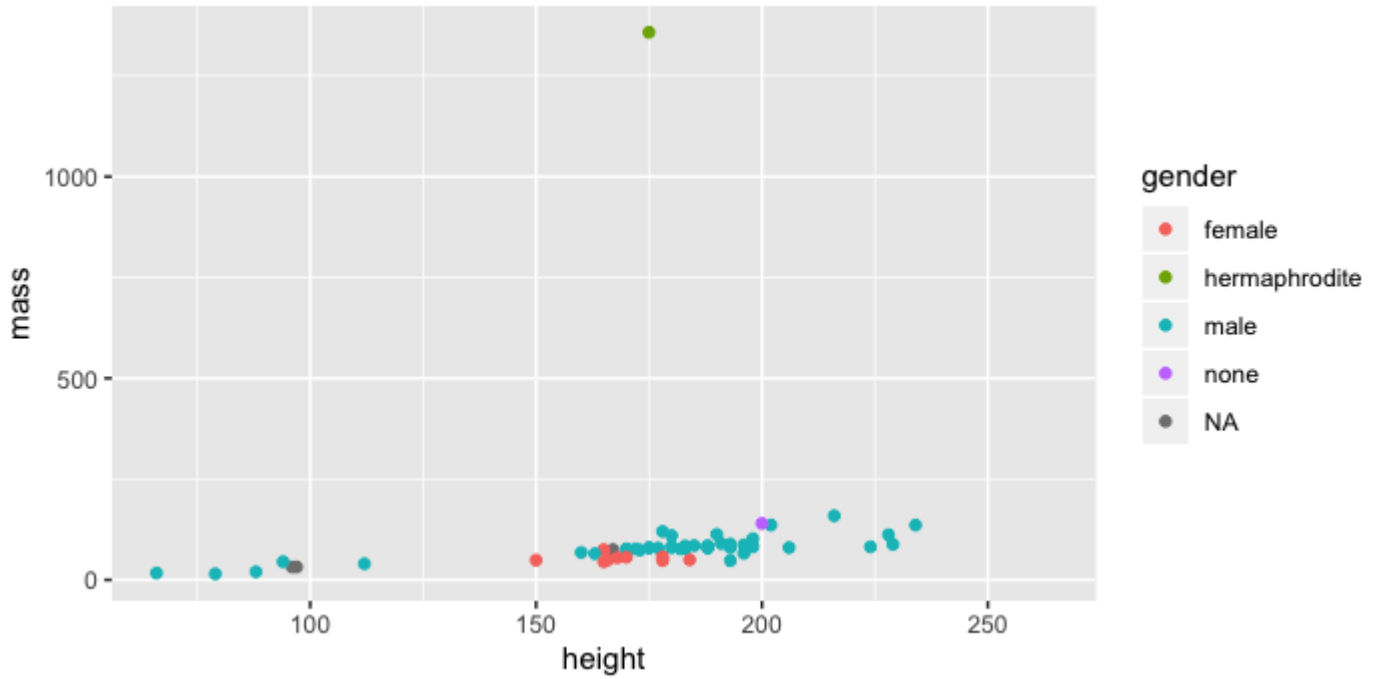
Aesthetics options

Visual characteristics of plotting characters that can be **mapped to a specific variable** in the data are

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

Mass vs. height + gender

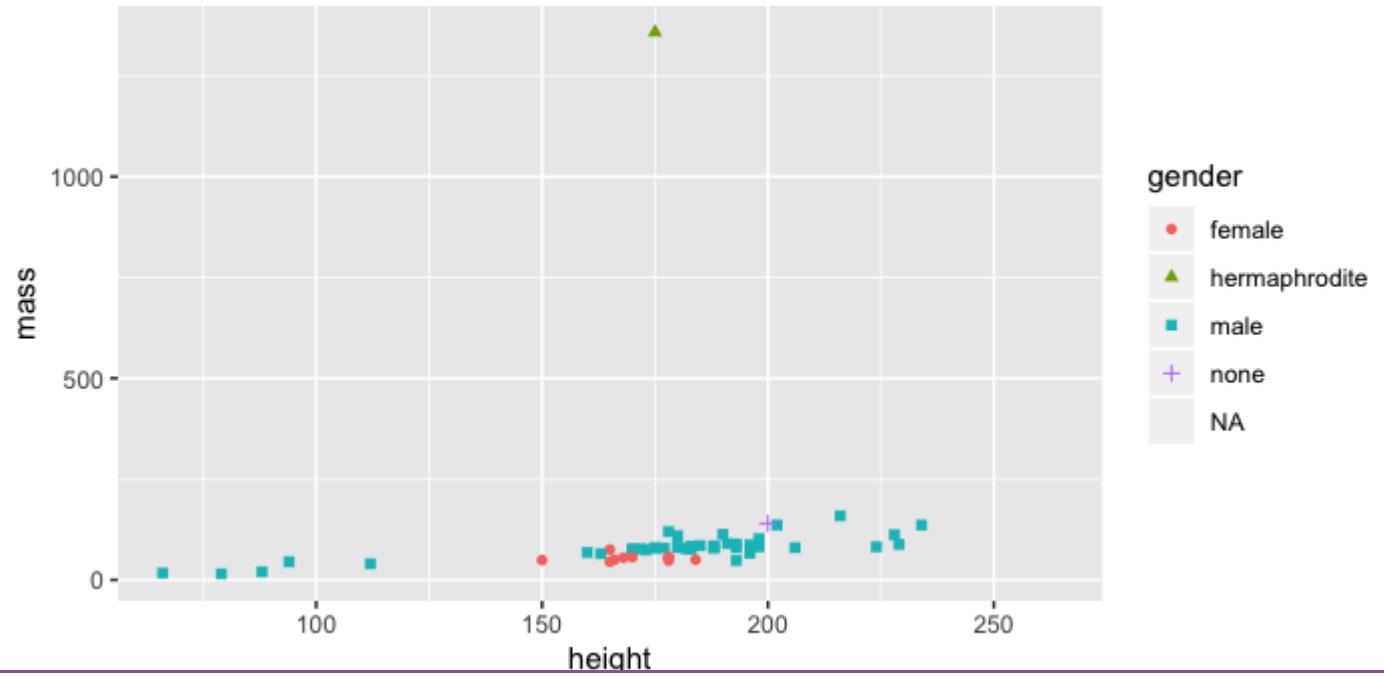
```
ggplot(data = starwars, mapping = aes(x = height, y = mass, color = gender))
  geom_point()
```



Mass vs. height + gender

Let's map **shape** and **color** to gender

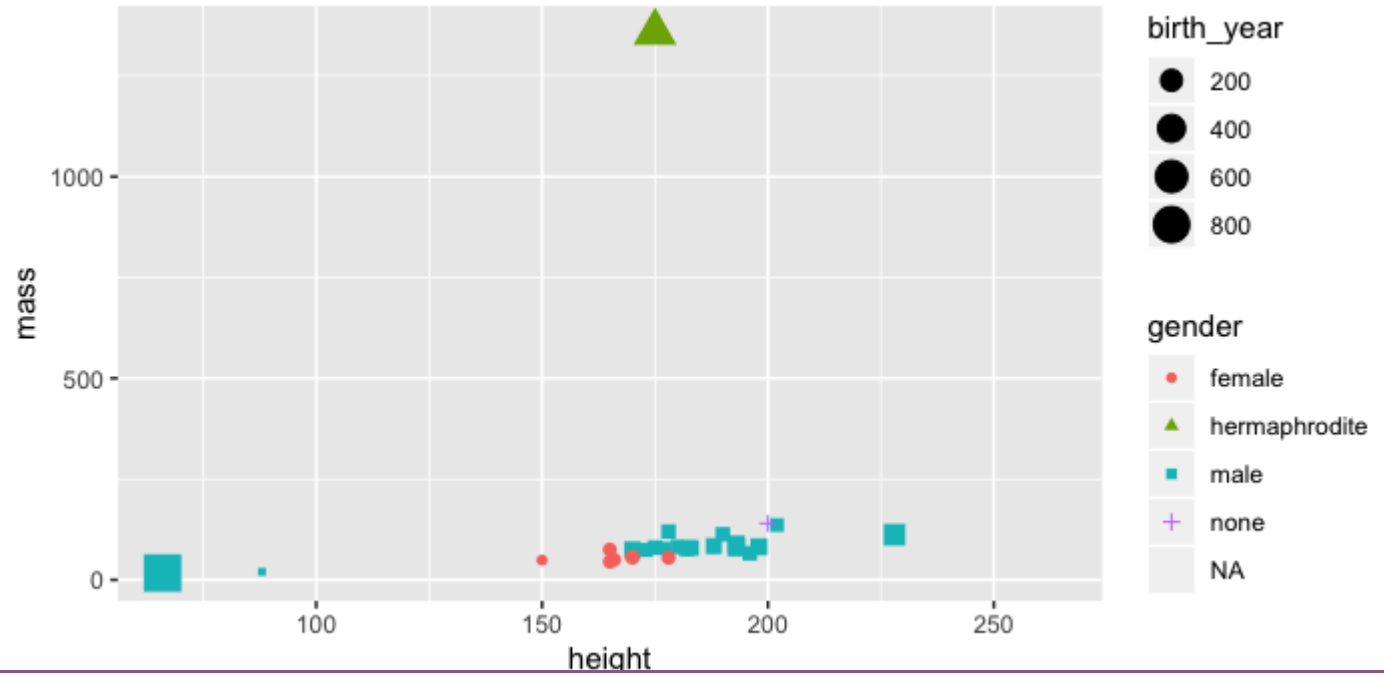
```
ggplot(data = starwars, mapping = aes(x = height, y = mass, color = gender,
                                      shape = gender
                                      )) +
  geom_point()
```



Mass vs. height + gender + birth year

Let's map the size to birth_year:

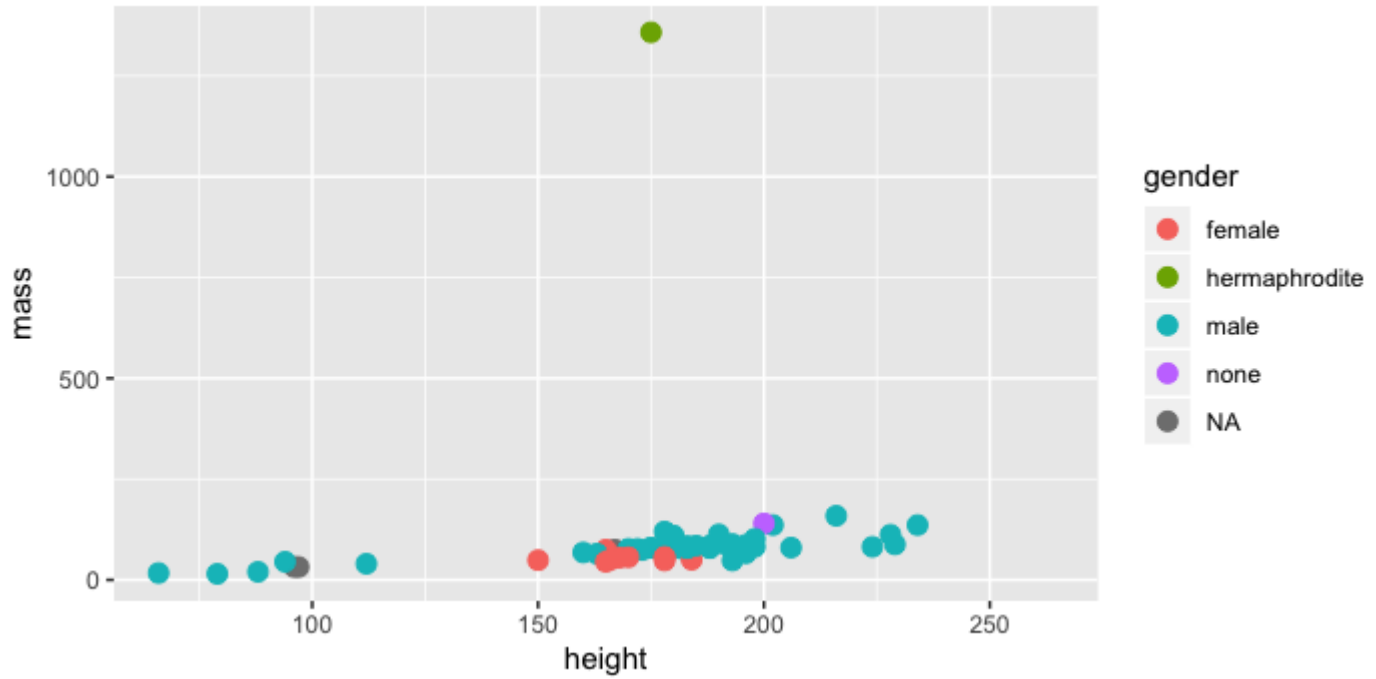
```
ggplot(data = starwars, mapping = aes(x = height, y = mass, color = gender, s
                                      size = birth_year
                                      )) +
  geom_point()
```



Mass vs. height + gender

Let's increase the size of all points **not** based on the values of a variable in the data:

```
ggplot(data = starwars, mapping = aes(x = height, y = mass, color = gender))
  geom_point(size = 3)
```



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

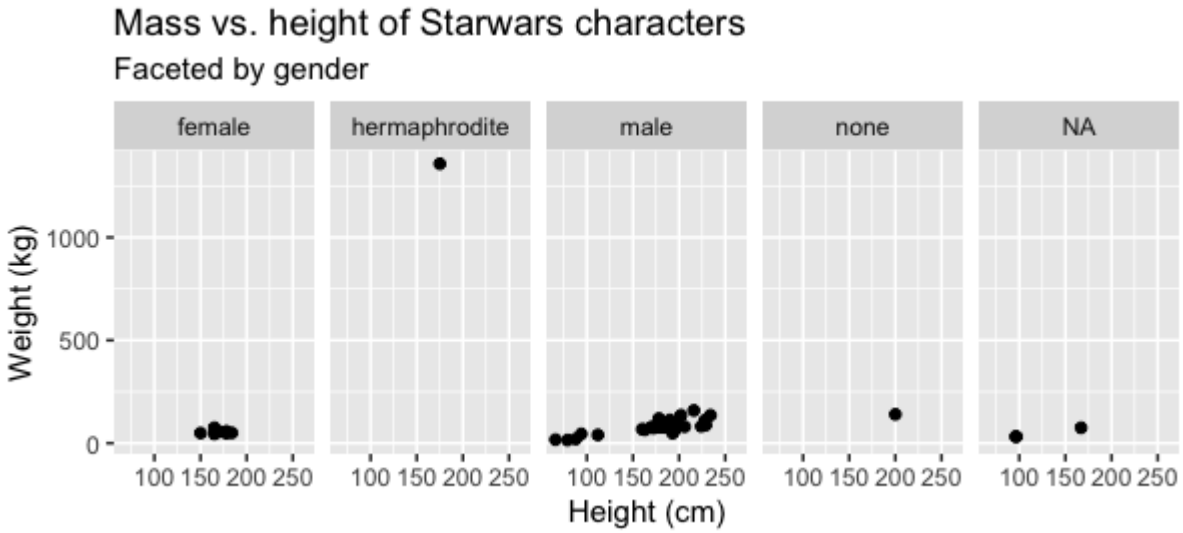
- Use aesthetics (**`aes`**) for mapping features of a plot to a variable, define the features in the **`geom_xxx`** for customization **not** mapped to a variable

Faceting

Faceting options

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

```
ggplot(data = starwars, mapping = aes(x = height, y = mass)) +
  facet_grid(. ~ gender) +
  geom_point() +
  labs(title = "Mass vs. height of Starwars characters",
       subtitle = "Faceted by gender",
       x = "Height (cm)", y = "Weight (kg)")
```

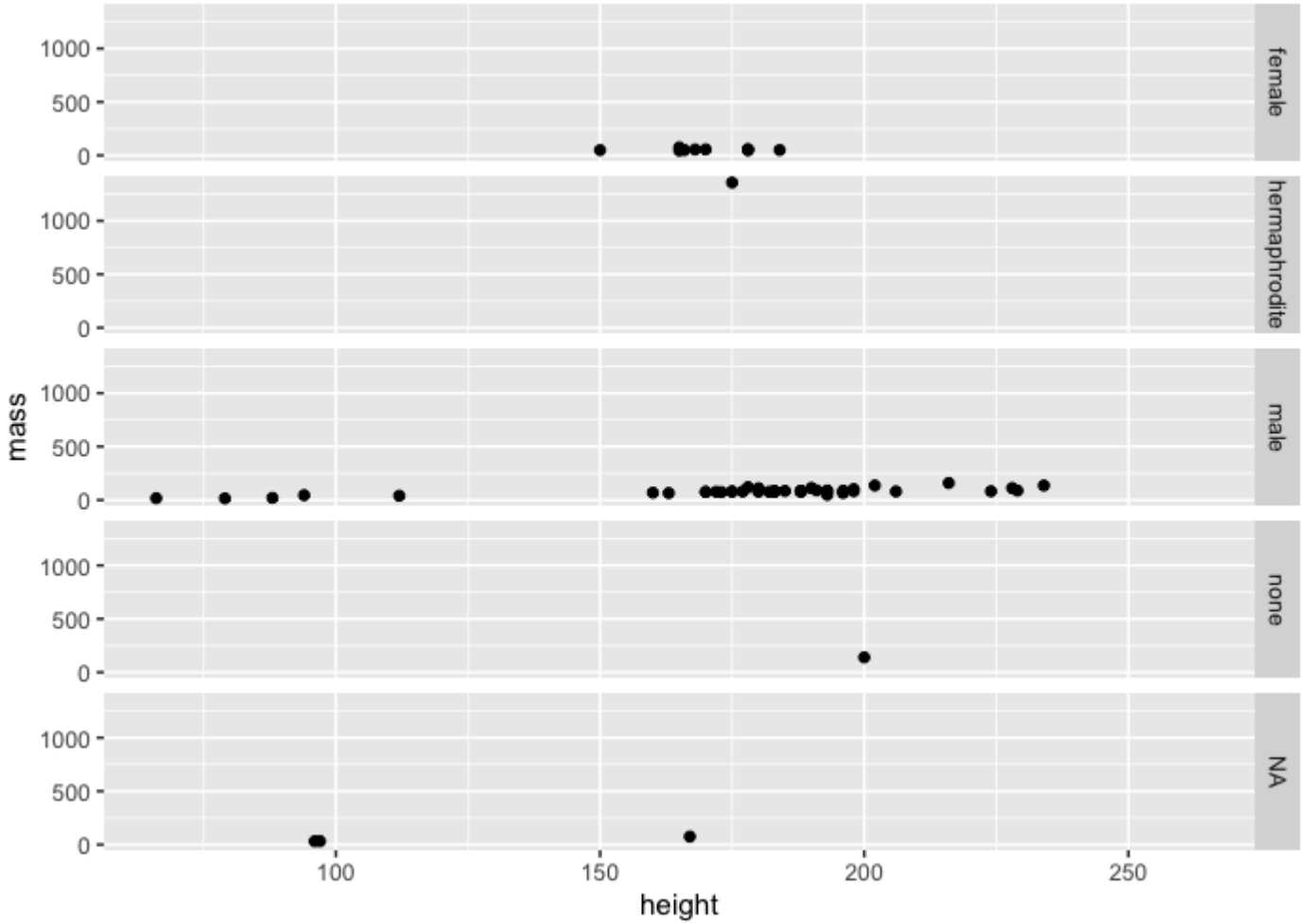


Dive further...

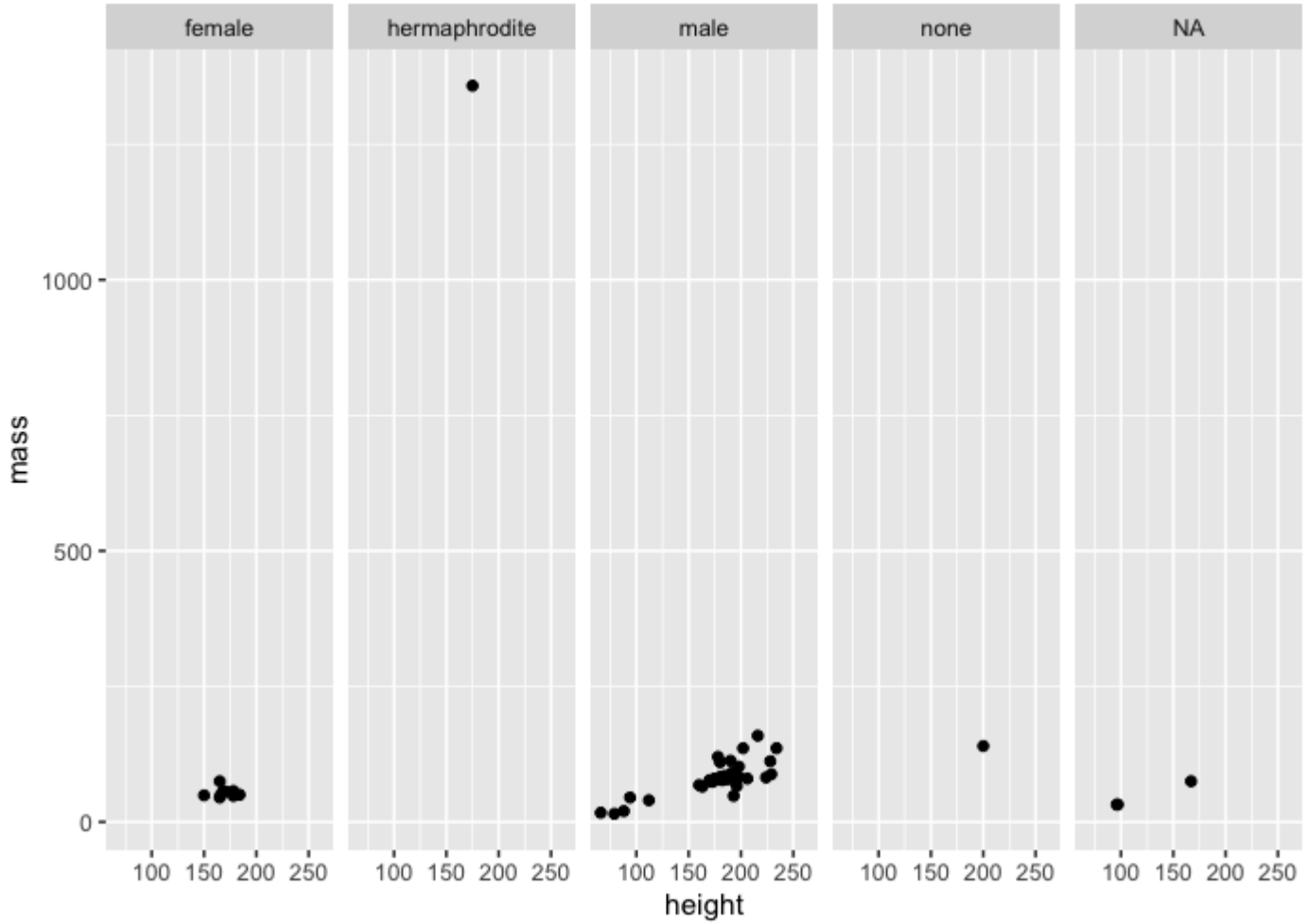
In the next few slides describe what each plot displays. Think about how the code relates to the output.

The plots in the next few slides do not have proper titles, axis labels, etc. because we want you to figure out what's happening in the plots. But you should always label your plots!

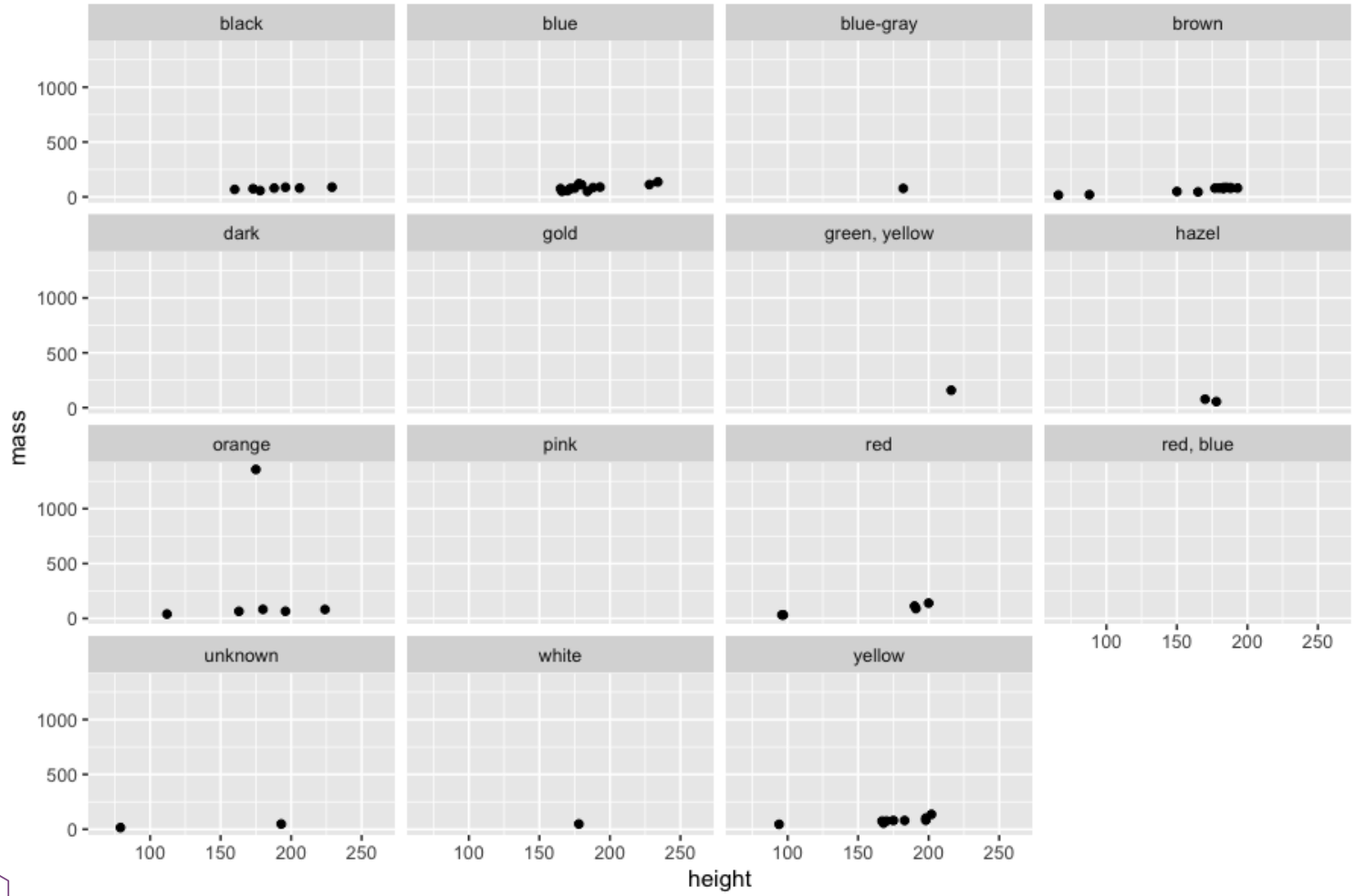
```
ggplot(data = starwars, mapping = aes(x = height, y = mass)) +
  geom_point() +
  facet_grid(gender ~ .)
```



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

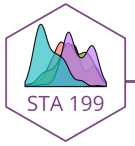


```
ggplot(data = starwars, mapping = aes(x = height, y = mass)) +
  geom_point() +
  facet_wrap(~ eye_color)
```



Facet summary

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



Starwars Application Exercise

- Go to https://github.com/sta199-fa19
- Click on the repo that begins with **`ae-03-starwars-`**.
- Click on **README.md** to see the instructions for this exercise.
- You will work in groups of 2 - 3 for the remainder of the exercise. One team member will go through the steps on their computer. The other team member(s) will follow along and read the instructions aloud from the README file.

Identifying variables

Number of variables involved

- Univariate data analysis: distribution of single variable
- Bivariate data analysis: relationship between two variables
- Multivariate data analysis: relationship between many variables at once, usually focusing on the relationship between two while conditioning for others

Types of variables

- Numerical variables can be classified as continuous or discrete based on whether or not the variable can take on an infinite number of values or only non-negative whole numbers, respectively.
  - *height* is continuous
  - *number of siblings* is discrete
- If the variable is categorical, we can determine if it is ordinal based on whether or not the levels have a natural ordering.
  - *hair color* is unordered
  - *year in school* is ordinal

Visualizing numerical data

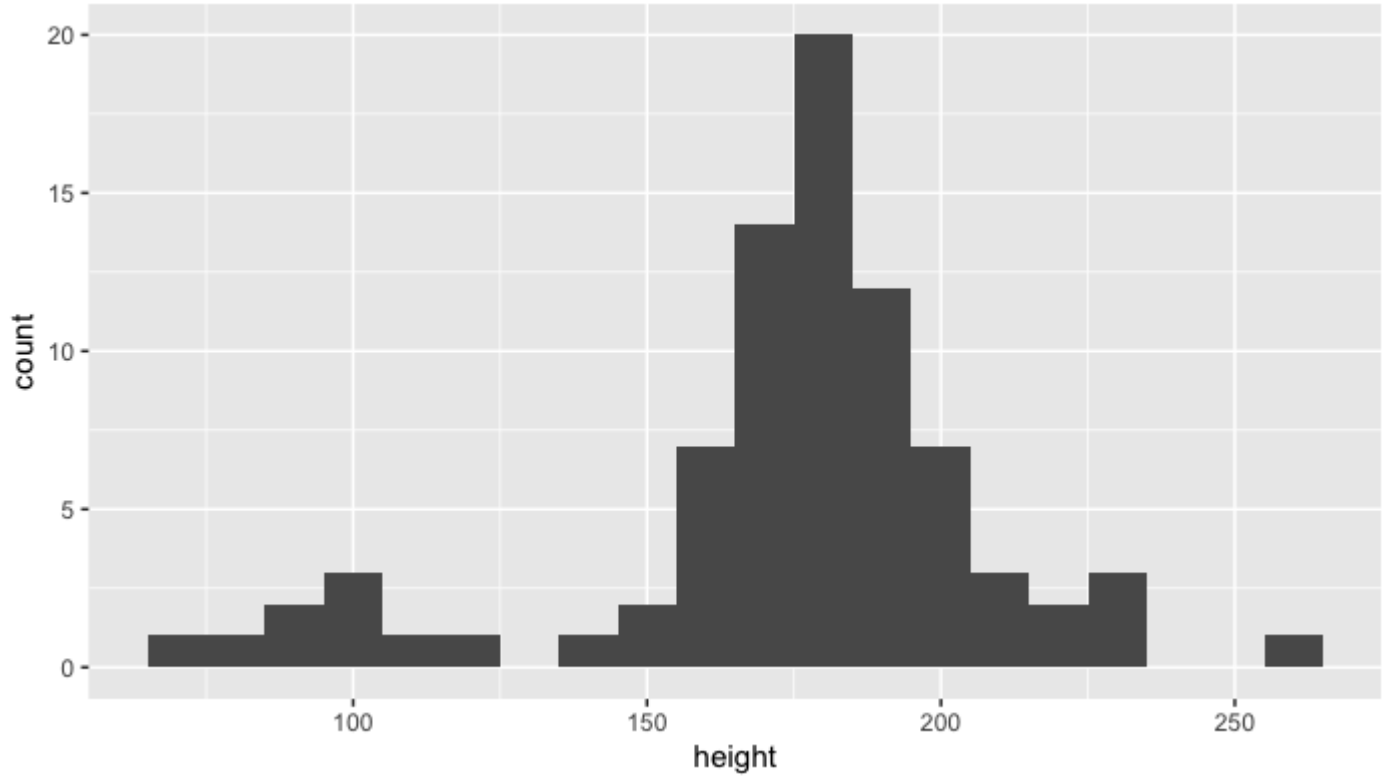
Describing shapes of numerical distributions

- **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`)
- **outliers**: observations outside of the usual pattern

Histograms

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

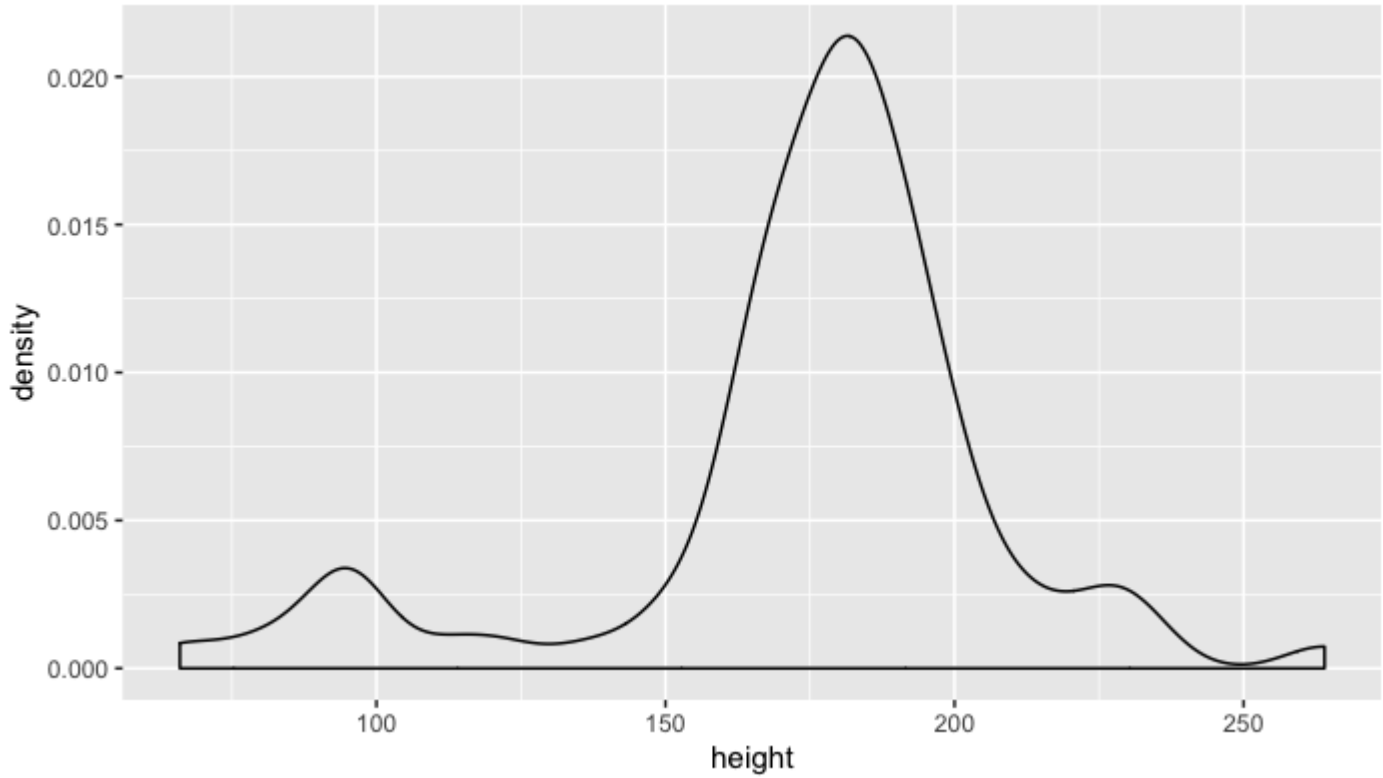
```
## Warning: Removed 6 rows containing non-finite values (stat_bin).
```



Density plots

```
ggplot(data = starwars, mapping = aes(x = height)) +
  geom_density()
```

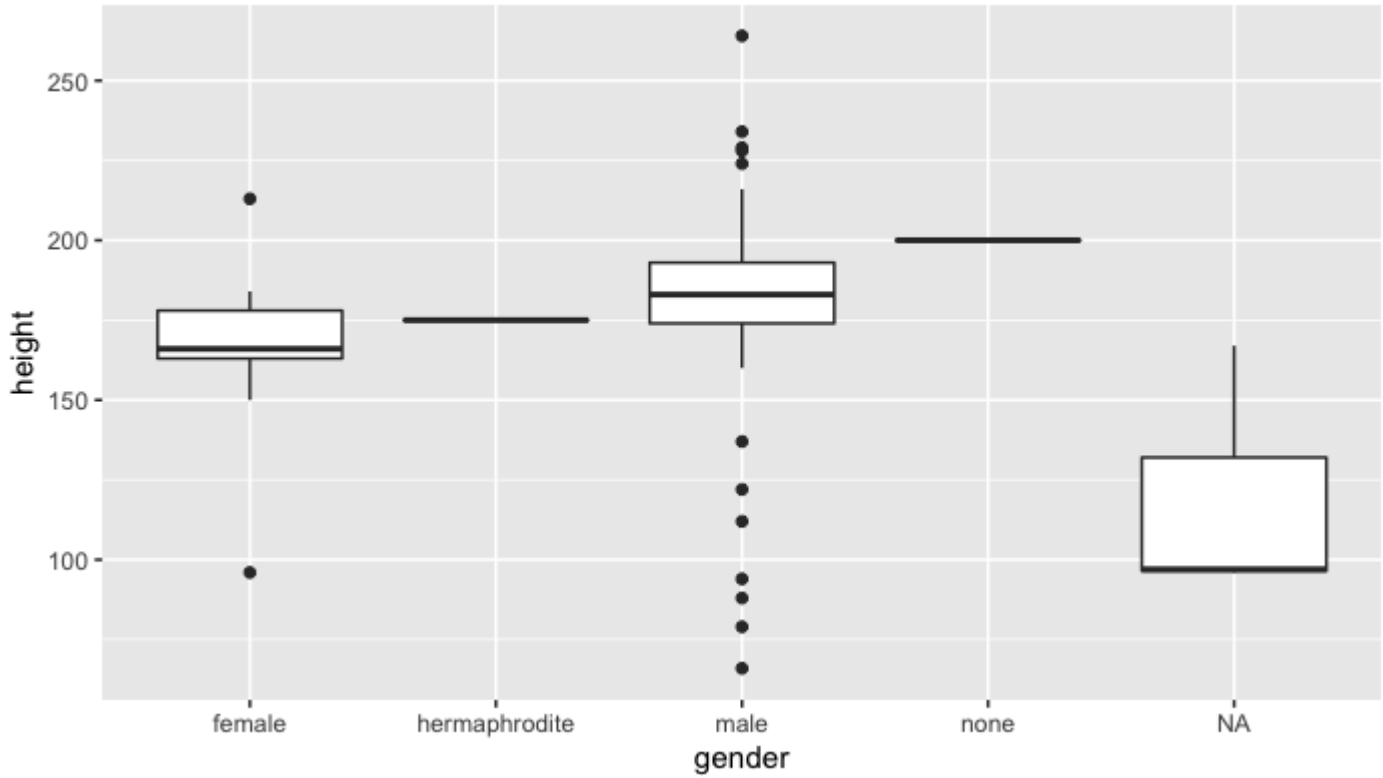
```
## Warning: Removed 6 rows containing non-finite values (stat_density).
```



Side-by-side box plots

```
ggplot(data = starwars, mapping = aes(y = height, x = gender)) +
  geom_boxplot()
```

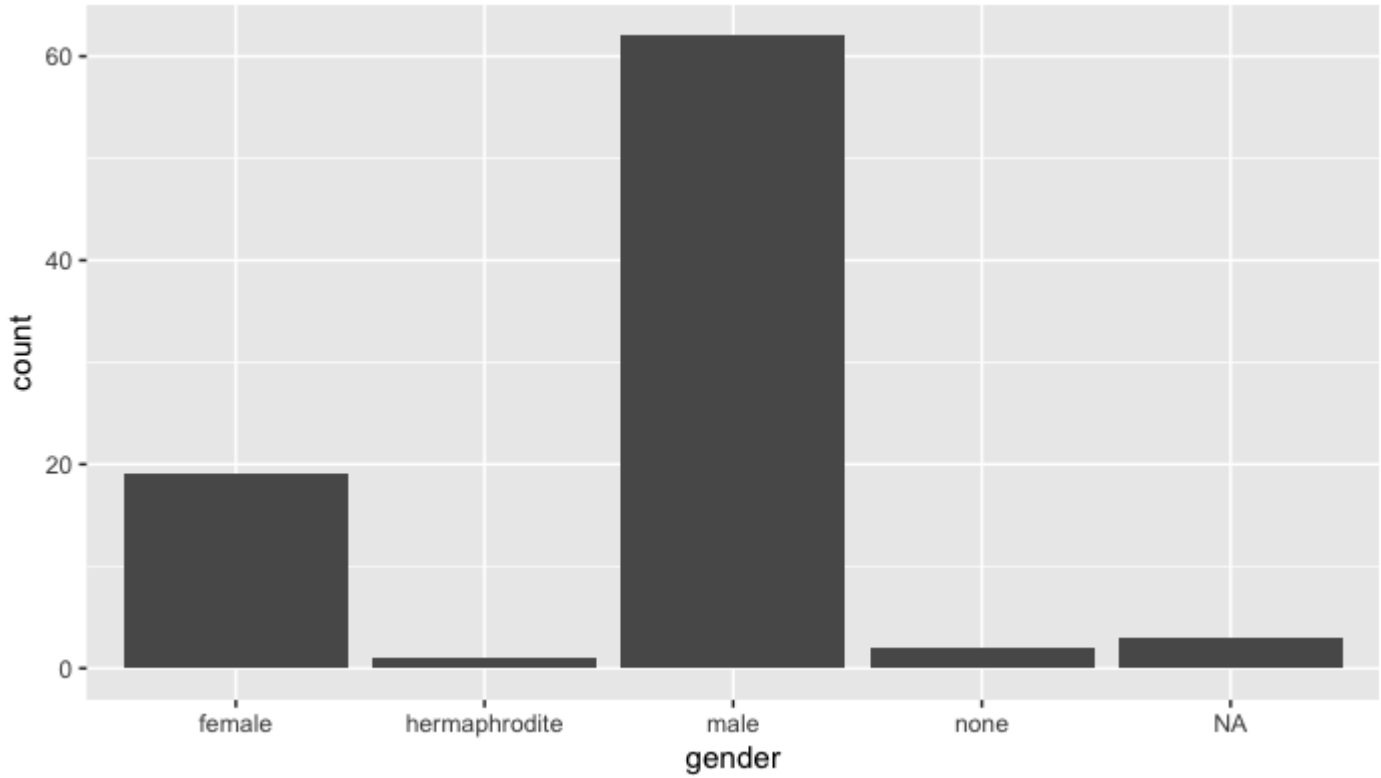
```
## Warning: Removed 6 rows containing non-finite values (stat_boxplot).
```



Visualizing categorical data

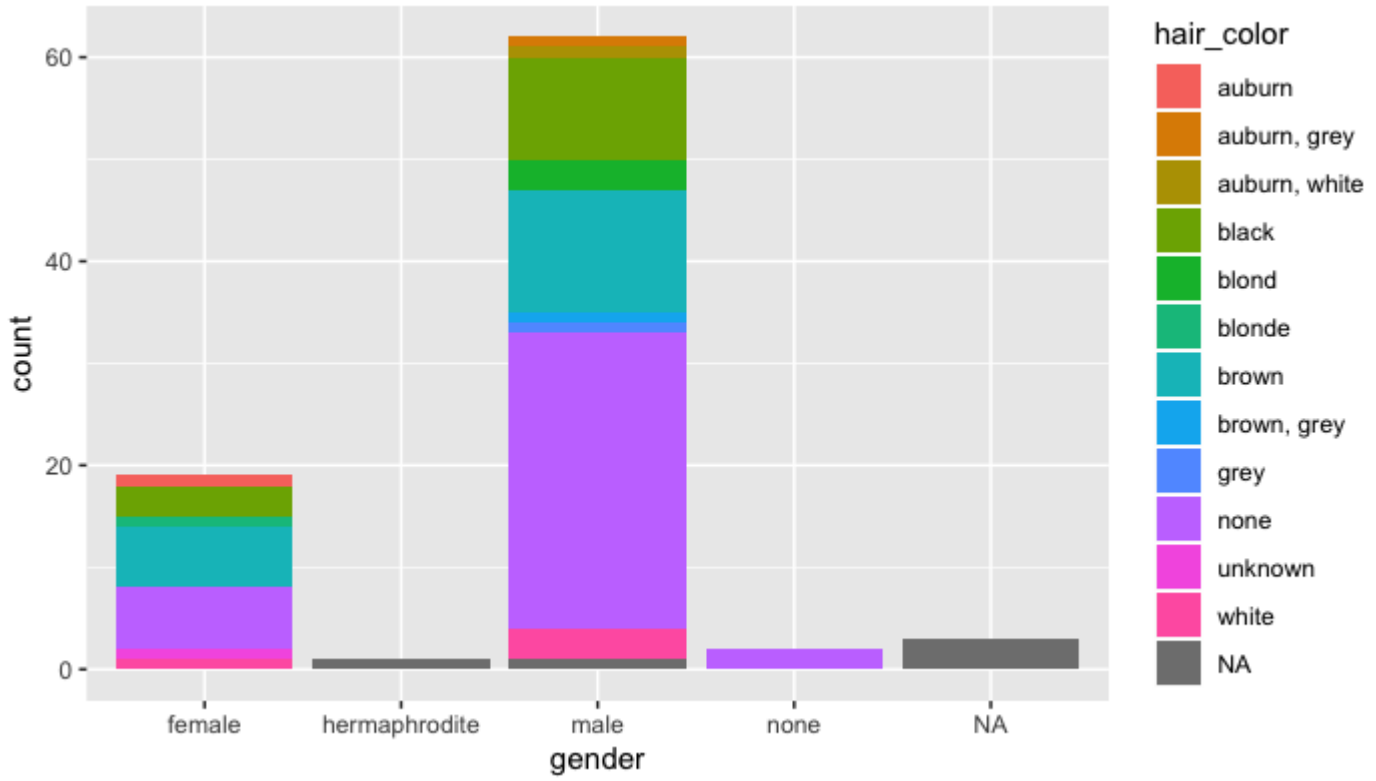
Bar plots

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



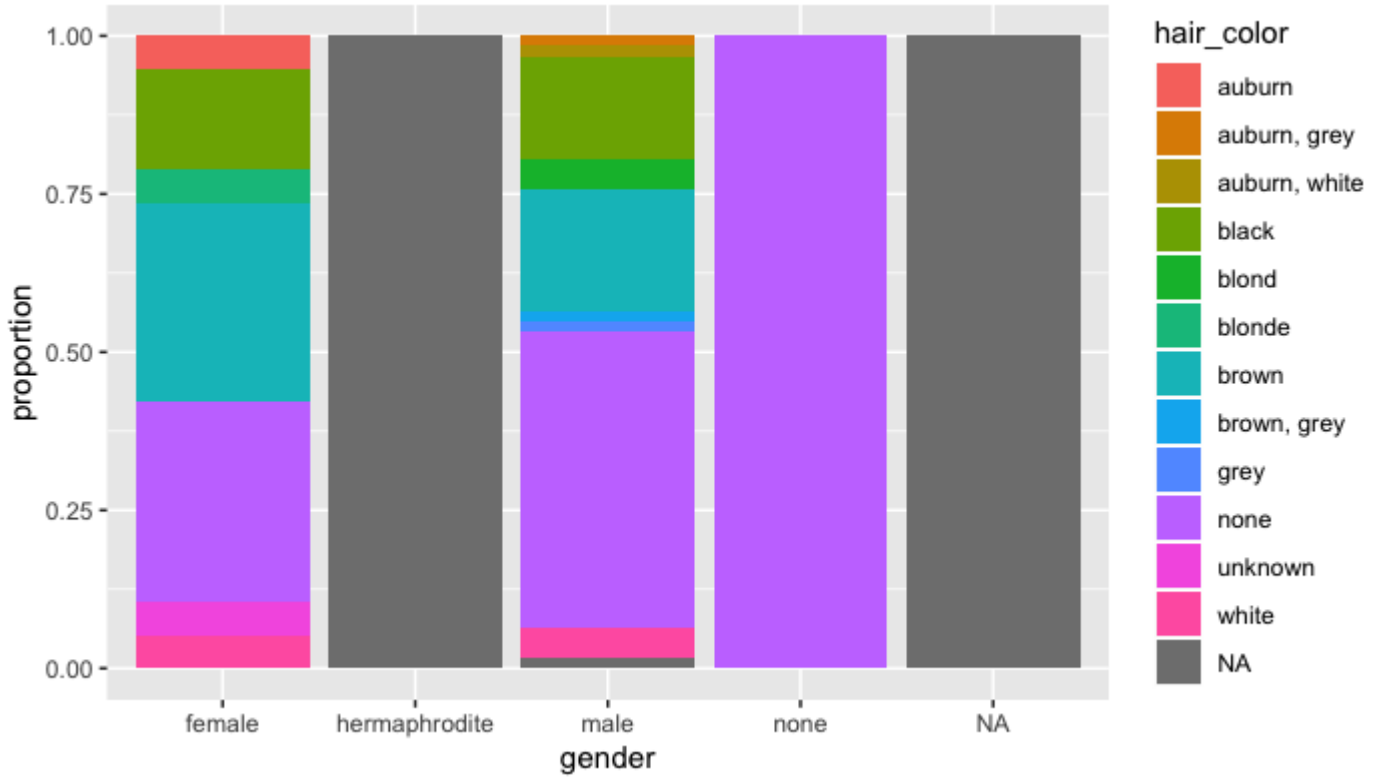
Segmented bar plots, counts

```
ggplot(data = starwars, mapping = aes(x = gender, fill = hair_color)) +
  geom_bar()
```



Segmented bar plots, proportions

```
ggplot(data = starwars, mapping = aes(x = gender, fill = hair_color)) +
  geom_bar(position = "fill") +
  labs(y = "proportion")
```



Which bar plot is more appropriate?

Which bar plot is a more useful representation for visualizing the relationship between gender and hair color? Why?

Before next class

- Start Reading 02 posted on the course schedule - due Thursday
- If you have not already done so,
  - complete "Getting to know you" survey on Sakai - due TODAY at 11:59p!
  - complete Lab 01 - due Thursday at 11:59p!