Visualization of star wars dataset using ggplot2

Lars Relund Nielsen

2020-11-10

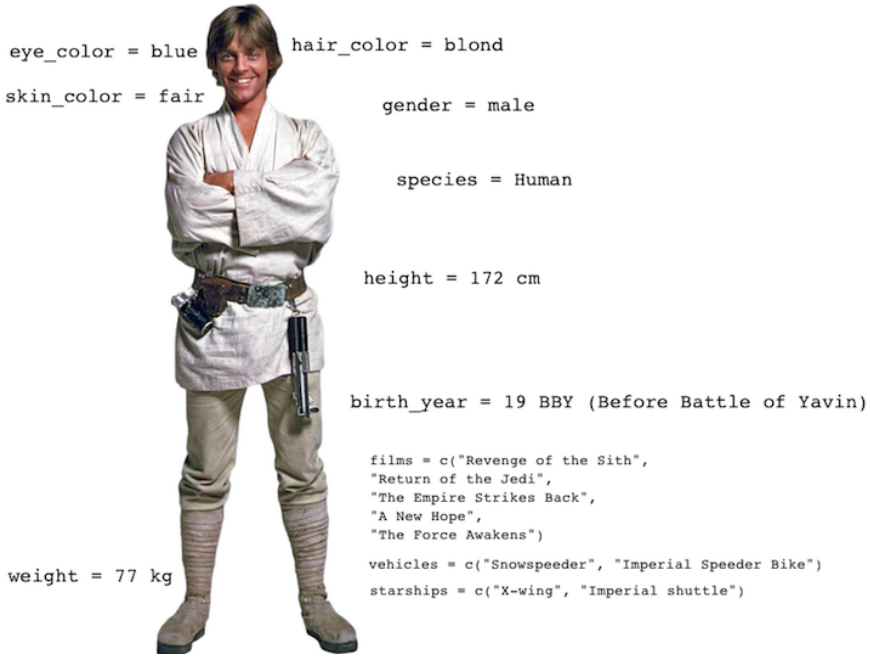
What's in the Star Wars data?

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

```
library(tidyverse)
# ?starwars
glimpse(starwars)
```

```
## Rows: 87
## Columns: 14
## $ name       <chr> "Luke Skywalker", "C-3PO", "R2-D2", "Darth Vader", "Leia Organa", "Owen Lars...
## $ height     <int> 172, 167, 96, 202, 150, 178, 165, 97, 183, 182, 188, 180, 228, 180, 173, 175...
## $ mass       <dbl> 77.0, 75.0, 32.0, 136.0, 49.0, 120.0, 75.0, 32.0, 84.0, 77.0, 84.0, NA, 112....
## $ hair_color <chr> "blond", NA, NA, "none", "brown", "brown, grey", "brown", NA, "black", "aubu...
## $ skin_color <chr> "fair", "gold", "white, blue", "white", "light", "light", "light", "white, r...
## $ eye_color  <chr> "blue", "yellow", "red", "yellow", "brown", "blue", "blue", "red", "brown", ...
## $ birth_year <dbl> 19.0, 112.0, 33.0, 41.9, 19.0, 52.0, 47.0, NA, 24.0, 57.0, 41.9, 64.0, 200.0...
## $ sex        <chr> "male", "none", "none", "male", "female", "male", "female", "none", "male", ...
## $ gender     <chr> "masculine", "masculine", "masculine", "masculine", "feminine", "masculine",...
## $ homeworld  <chr> "Tatooine", "Tatooine", "Naboo", "Tatooine", "Alderaan", "Tatooine", "Tatooi...
## $ species    <chr> "Human", "Droid", "Droid", "Human", "Human", "Human", "Human", "Droid", "Hum...
## $ films      <list> [<"The Empire Strikes Back", "Revenge of the Sith", "Return of the Jedi", "...
## $ vehicles   <list> [<"Snowspeeder", "Imperial Speeder Bike">, <>, <>, <>, "Imperial Speeder Bi...
## $ starships  <list> [<"X-wing", "Imperial shuttle">, <>, <>, "TIE Advanced x1", <>, <>, <>, <>,...
```

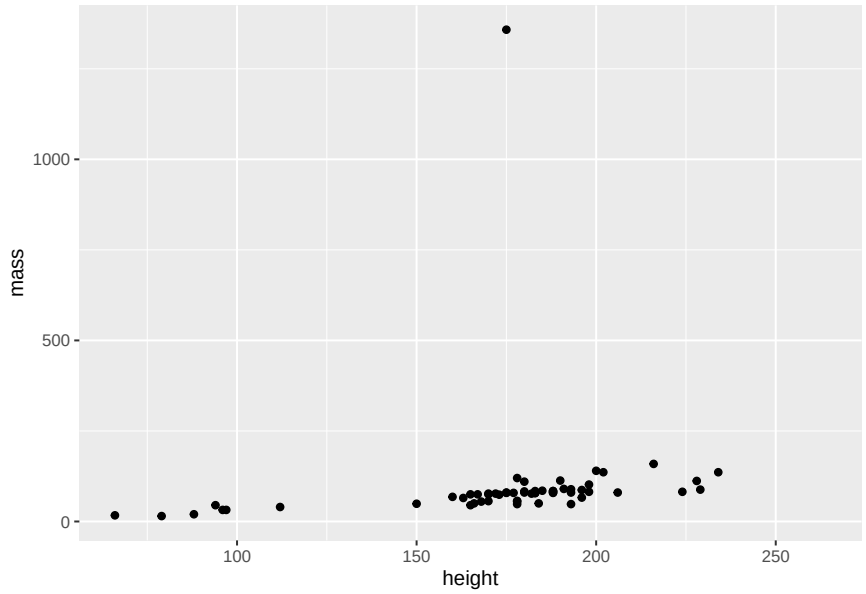
A row for each person in the movies. For instance the row for Luke Skywalker is:



Scatterplot of mass vs. height

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

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

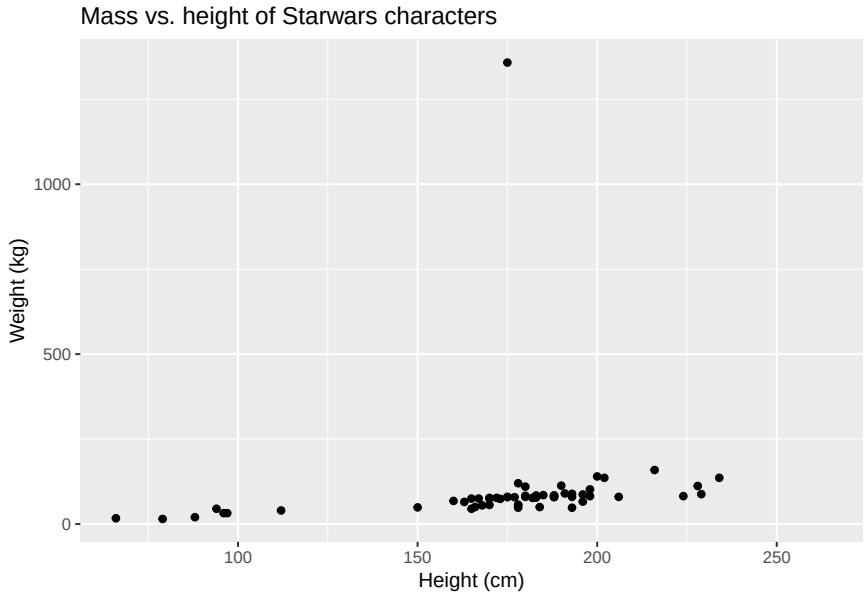


Note a warning is issued. Not all characters have height and mass information (hence 28 of them not plotted). You may exclude warnings in chunk output by using option `warning = FALSE`. However, but often it's important to note it.

Adding labels

The `labs` function can be used to add different labels to the plot.

```
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)")
```



The fat guy

- 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 chubby character?

```
starwars %>%
  filter(mass > 500) %>%
  select(name, height, mass)
```

```
## # A tibble: 1 x 3
##   name                  height  mass
##   <chr>                  <int> <dbl>
## 1 Jabba Desilijic Tiure    175  1358
```

Aesthetics

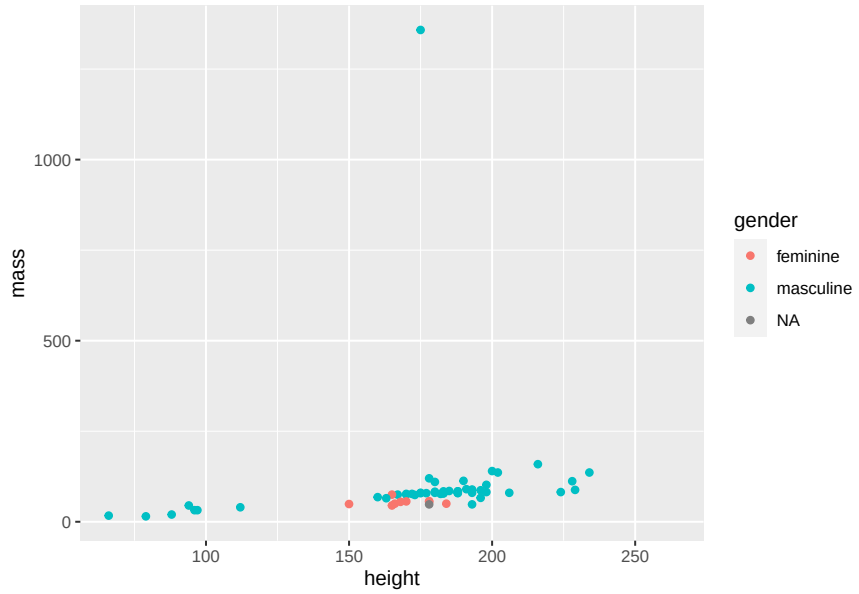
Visual characteristics of plotting can be **mapped to a specific variable** using aesthetics. For instance, the visual characteristics of plotting points are:

- `color`
- `size`

- `shape`
- `alpha` (transparency)

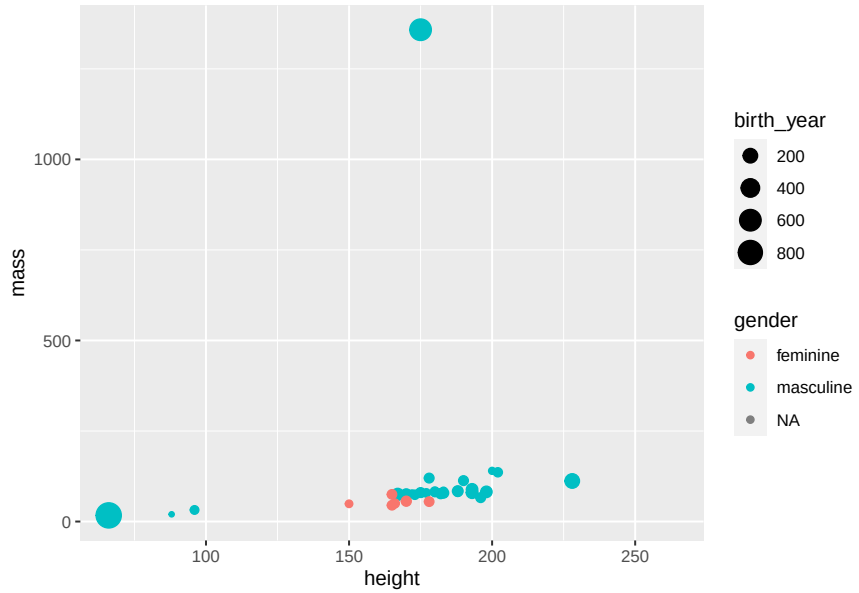
Let us have a look at mass vs. height using gender as color aesthetic:

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



In general not many girls in star wars. Let us add birth year as size aesthetic:

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

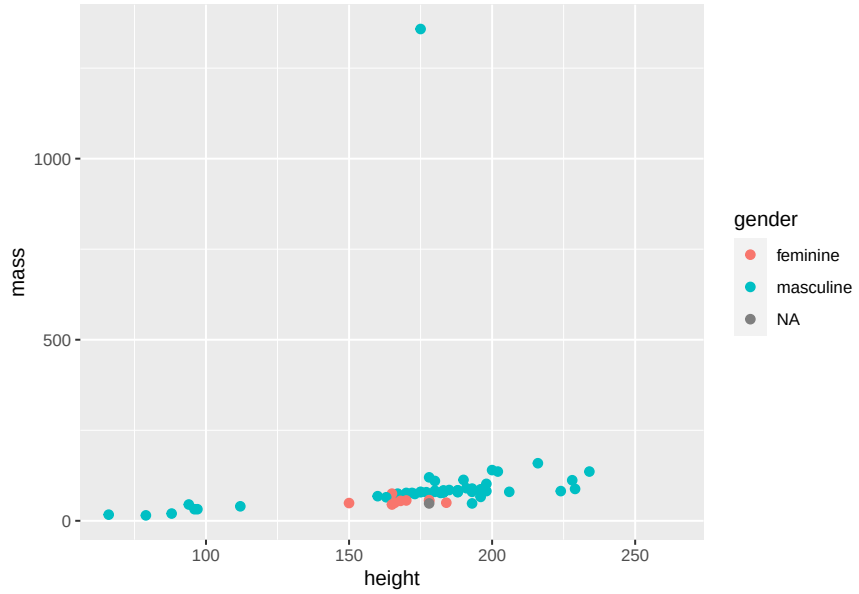


Jabba is very old. What about species? We use species as color aesthetic:

```
# Your turn - Finish the code
 ggplot(data = starwars,
        mapping = aes(height, mass, color = species)) +
   geom_point()
```

Let's now 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 = 2)
```



A plot of mass given height with hair color as color aesthetic and using fixed `shape = 0`:

```
# Your turn: finish the code
 ggplot(data = starwars, mapping = aes(x = height, y = mass, color = hair_color)) +
   geom_point(shape = 0)
```

Summary

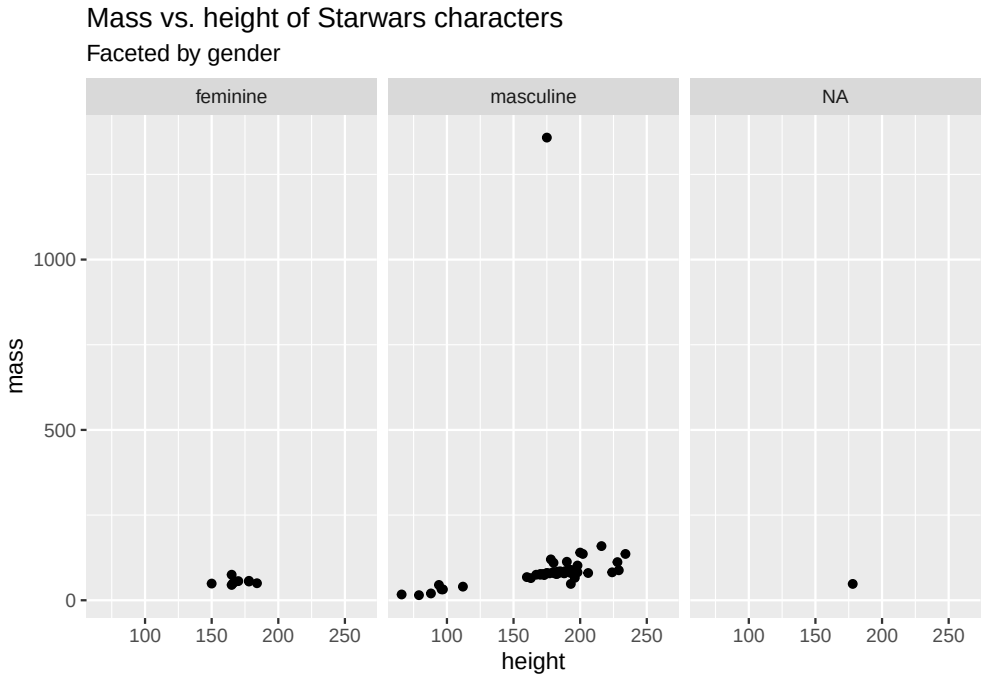
- Continuous variables are measured on a continuous scale.
- Discrete variables are measured (or often counted) on a discrete scale.
- Use aesthetics for mapping features of a plot to a variable.
- Define fixed features in the geom outside the `aes` function.

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

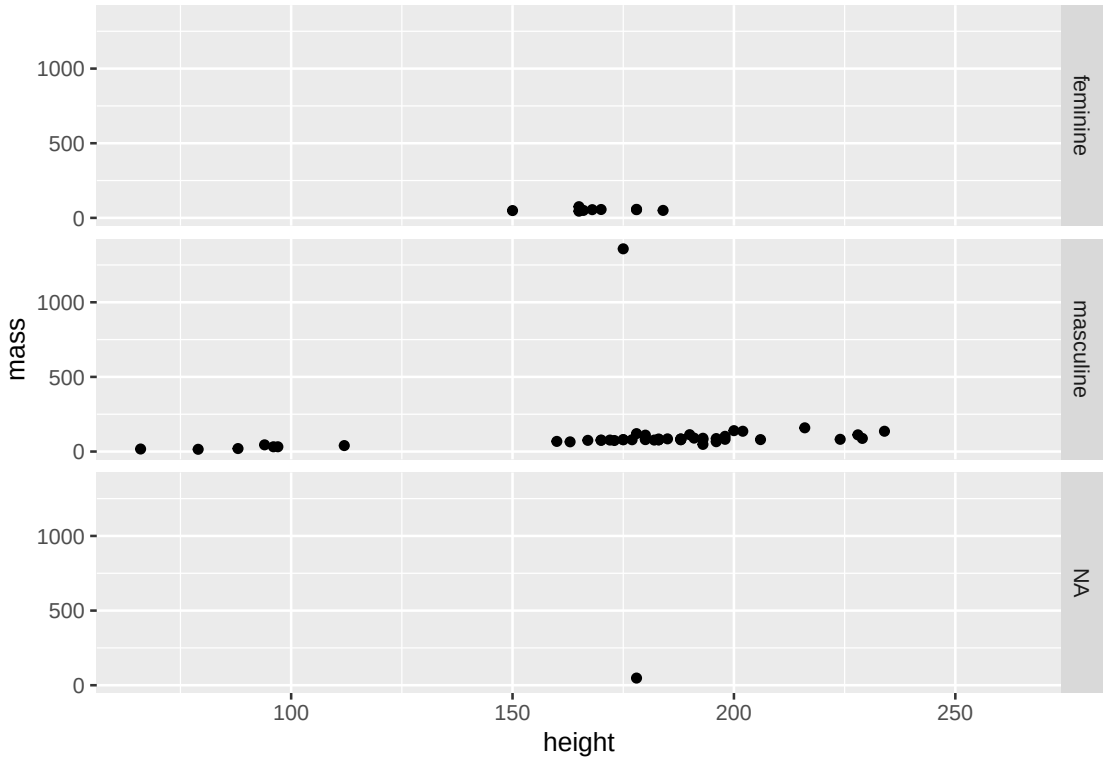
- Smaller plots that display different subsets of the data.
- Useful for exploring conditional relationships and large data.

```
ggplot(data = starwars, mapping = aes(x = height, y = mass)) +
  facet_grid(cols = vars(gender)) +  # or use rows/cols argument
  geom_point() +
  labs(title = "Mass vs. height of Starwars characters",
       subtitle = "Faceted by gender")
```

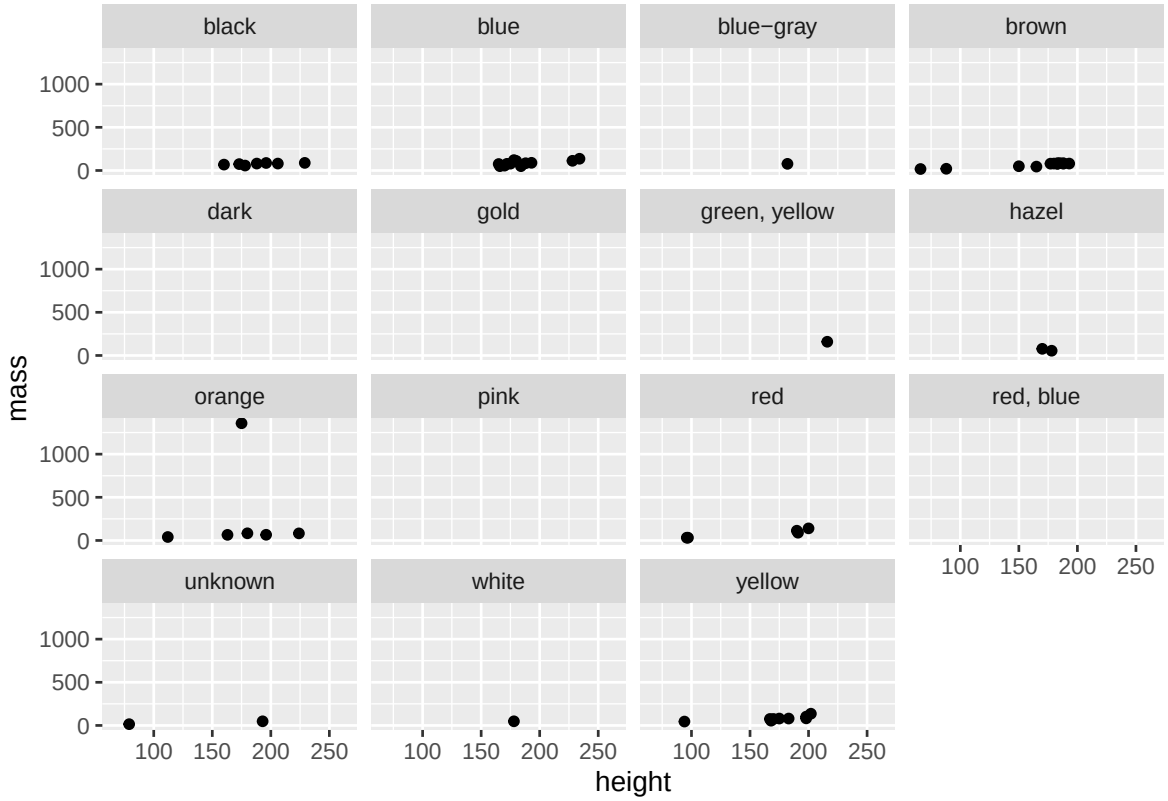


- Describe what each plot displays.
- Think about how the code relates to the output.

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



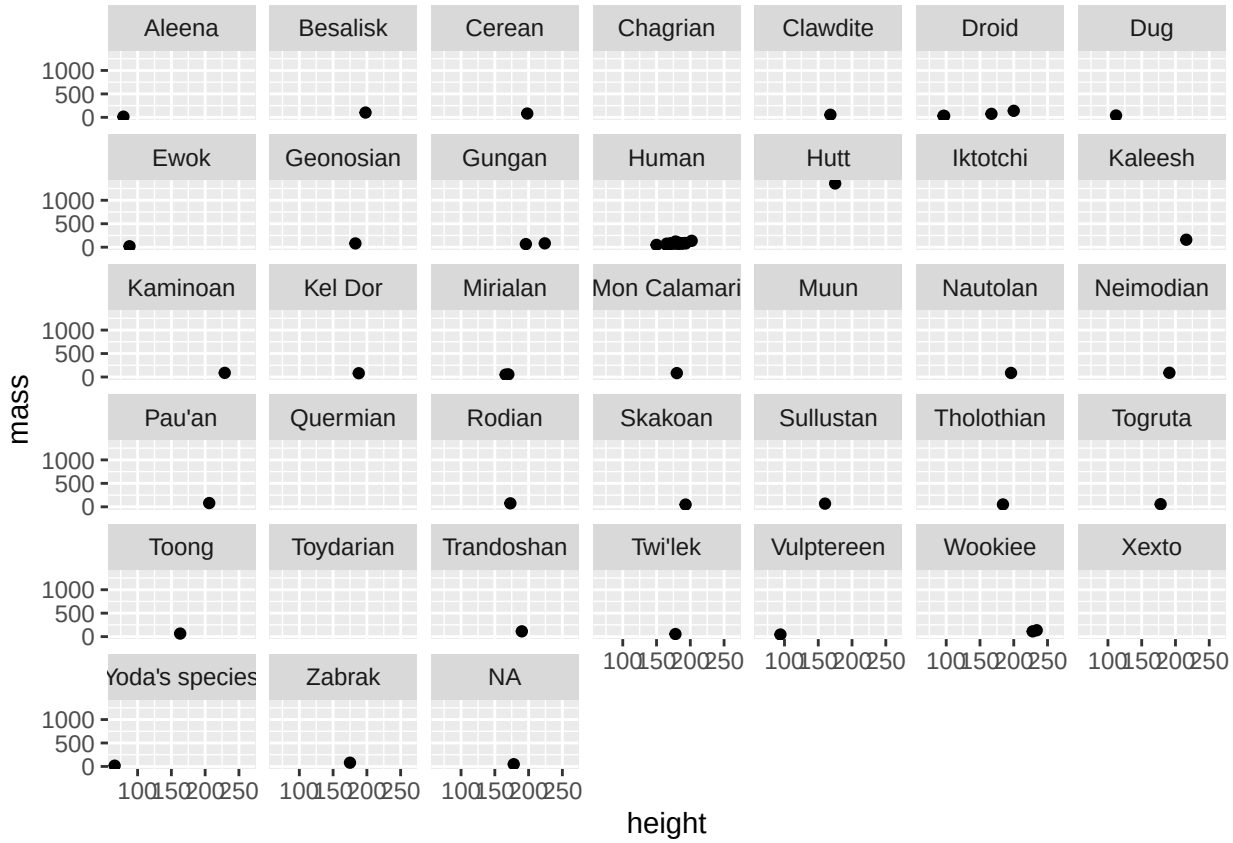

```
ggplot(data = starwars, mapping = aes(x = height, y = mass)) +
  geom_point() +
  facet_wrap(vars(eye_color))
```



Let us facet species:

```
# Your turn - Finish the code
ggplot(data = starwars, mapping = aes(x = height, y = mass)) +
  geom_point() +
  facet_wrap(vars(species))
```

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



Summary

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

Identifying variables and plot type

- 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

There are different variable types:

- **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.
- If the variable is **categorical**, we can determine if it is **ordinal** based on whether or not the levels have a natural ordering.

Visualizing numerical data

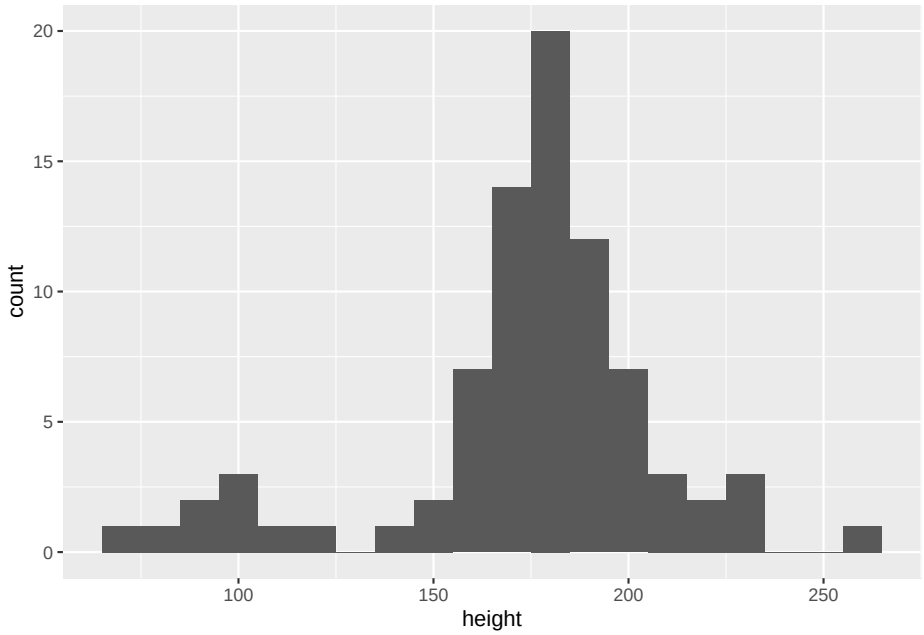
Describing shapes of numerical distributions

- 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`)
- unusal observations

Histograms

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

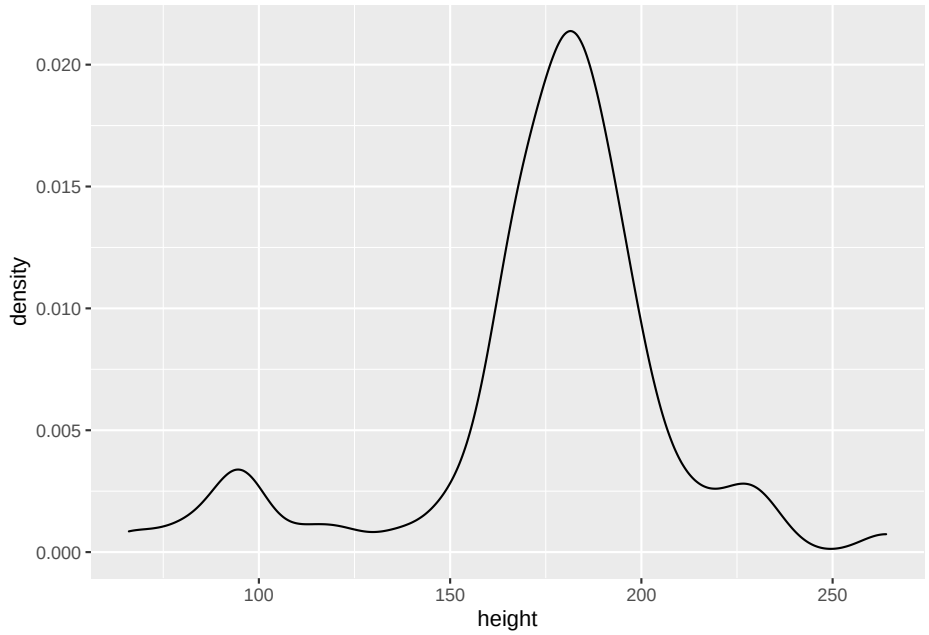
```
## Warning: Removed 6 rows containing non-finite values (stat_bin).
```



Density plots

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

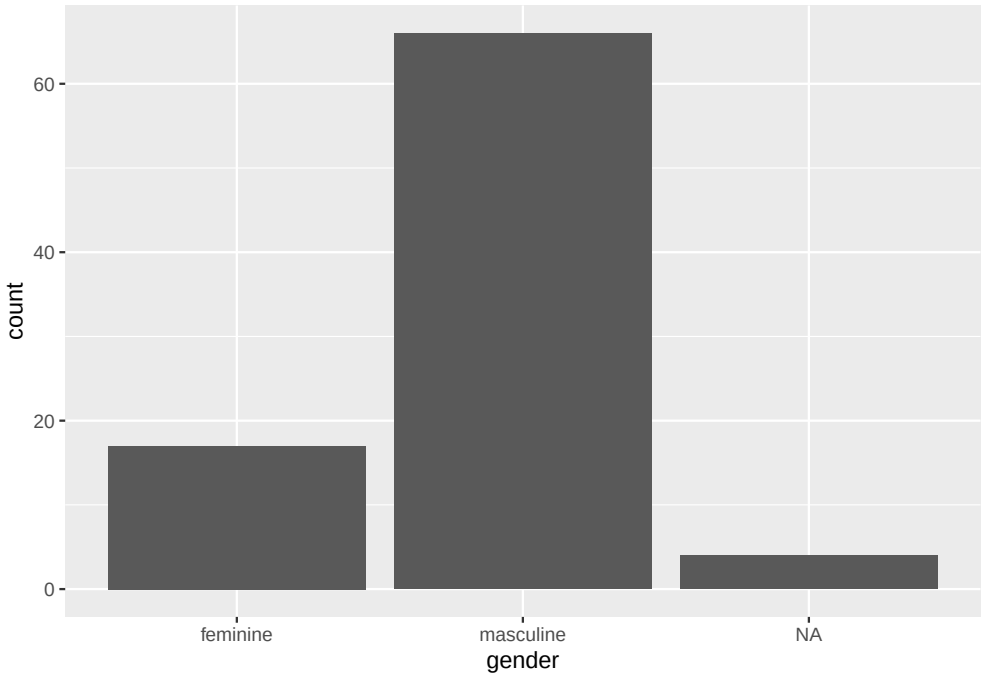
```
## Warning: Removed 6 rows containing non-finite values (stat_density).
```



Visualizing categorical data

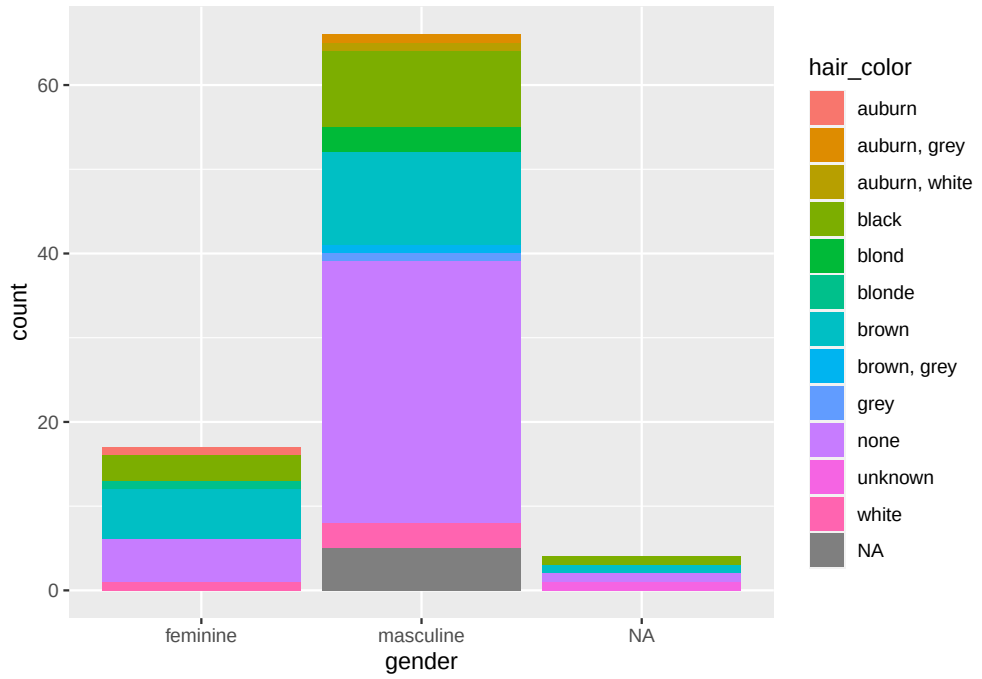
Bar plots

```
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()
```

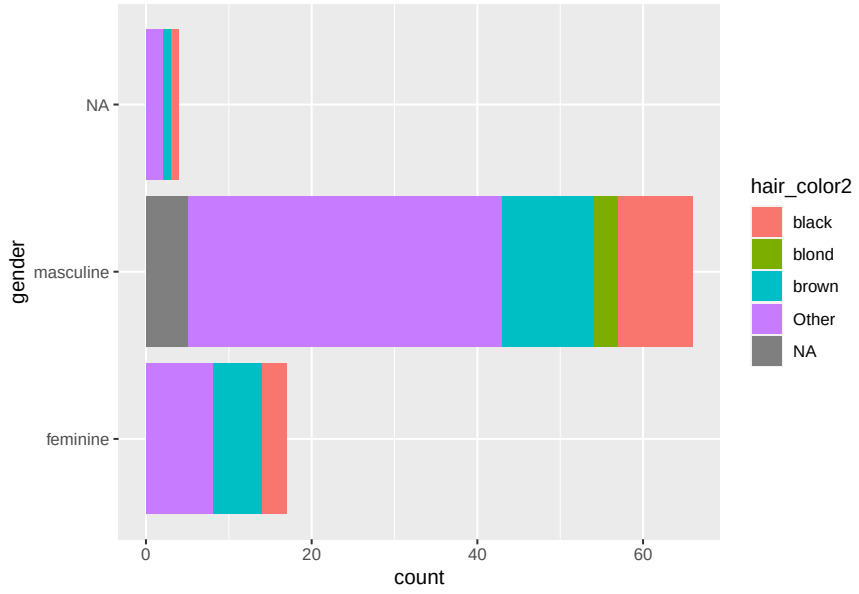


Let us recode hair color into a smaller set:

```
starwars <- starwars %>%
  mutate(hair_color2 =
           fct_other(hair_color,
                     keep = c("black", "brown", "blond")
                     )
         )
```

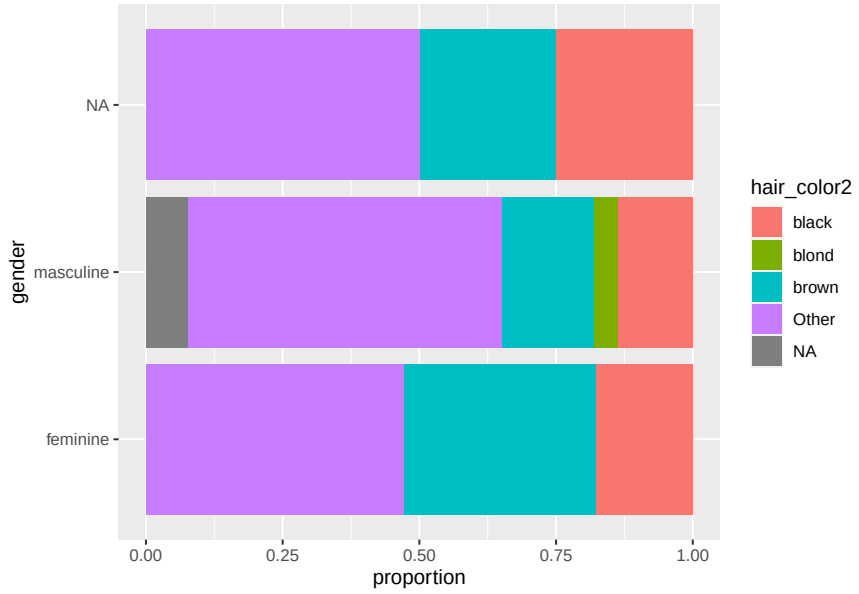
The segmented bar plot now becomes

```
# Your turn - Finish the code
ggplot(data = starwars,
       mapping = aes(x = gender, fill = hair_color2)) +
  geom_bar() +
  coord_flip()
```



Try to run the following:

```
ggplot(data = starwars,
       mapping = aes(x = gender, fill = hair_color2)) +
  geom_bar(position = "fill") +
  coord_flip() +
  labs(y = "proportion")
```

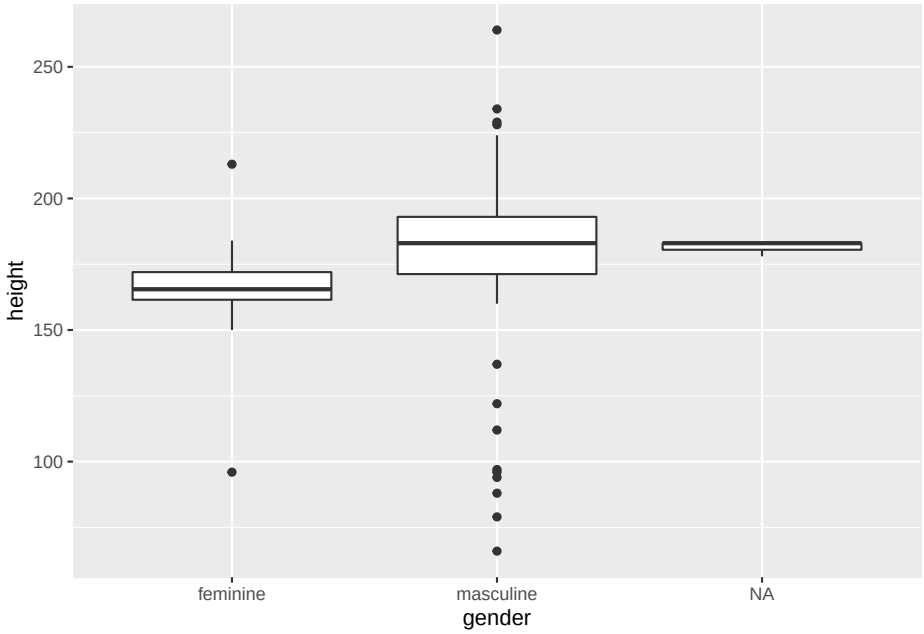


Which bar plot is a more useful representation for visualizing the relationship between gender and hair color?

Visualizing relationships between numerical and categorical data

Side-by-side box plots

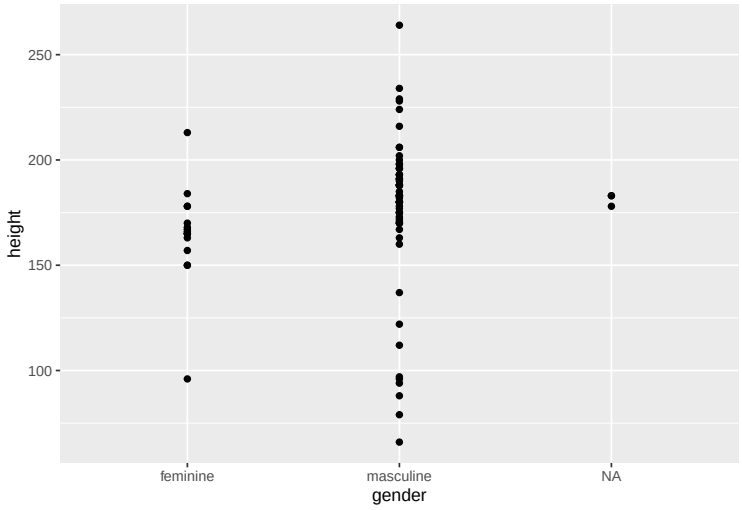
```
ggplot(data = starwars, mapping = aes(y = height, x = gender)) +
  geom_boxplot()
```



Scatter plot...

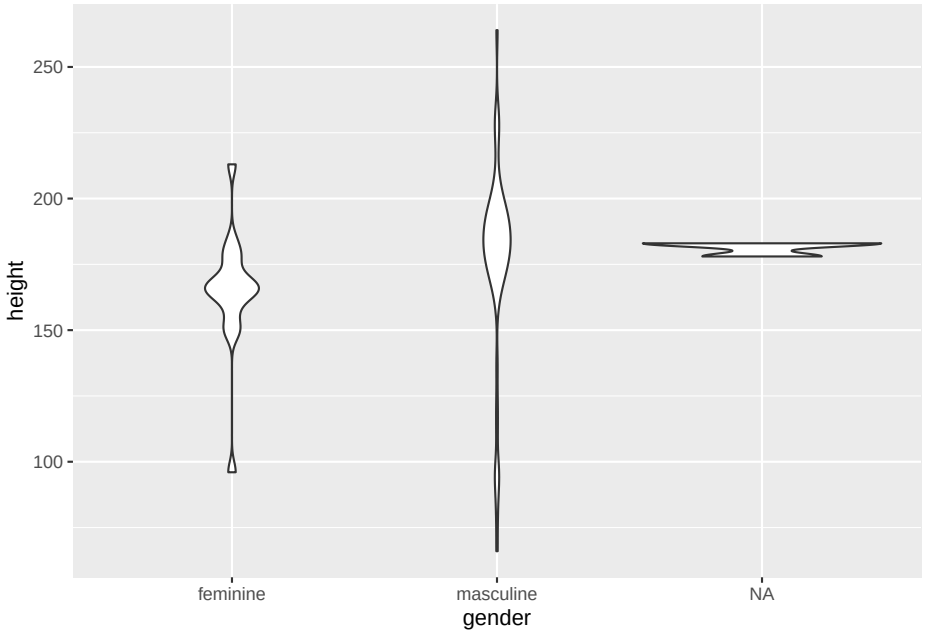
This is not a great representation of these data.

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



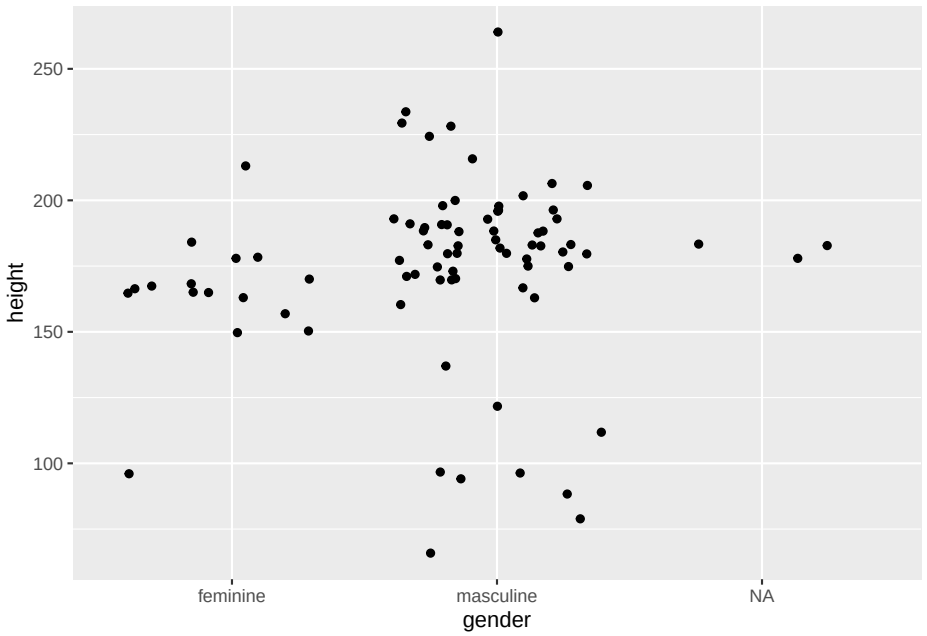
Violin plots

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



Jitter plot

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



What does `geom_gitter` do?

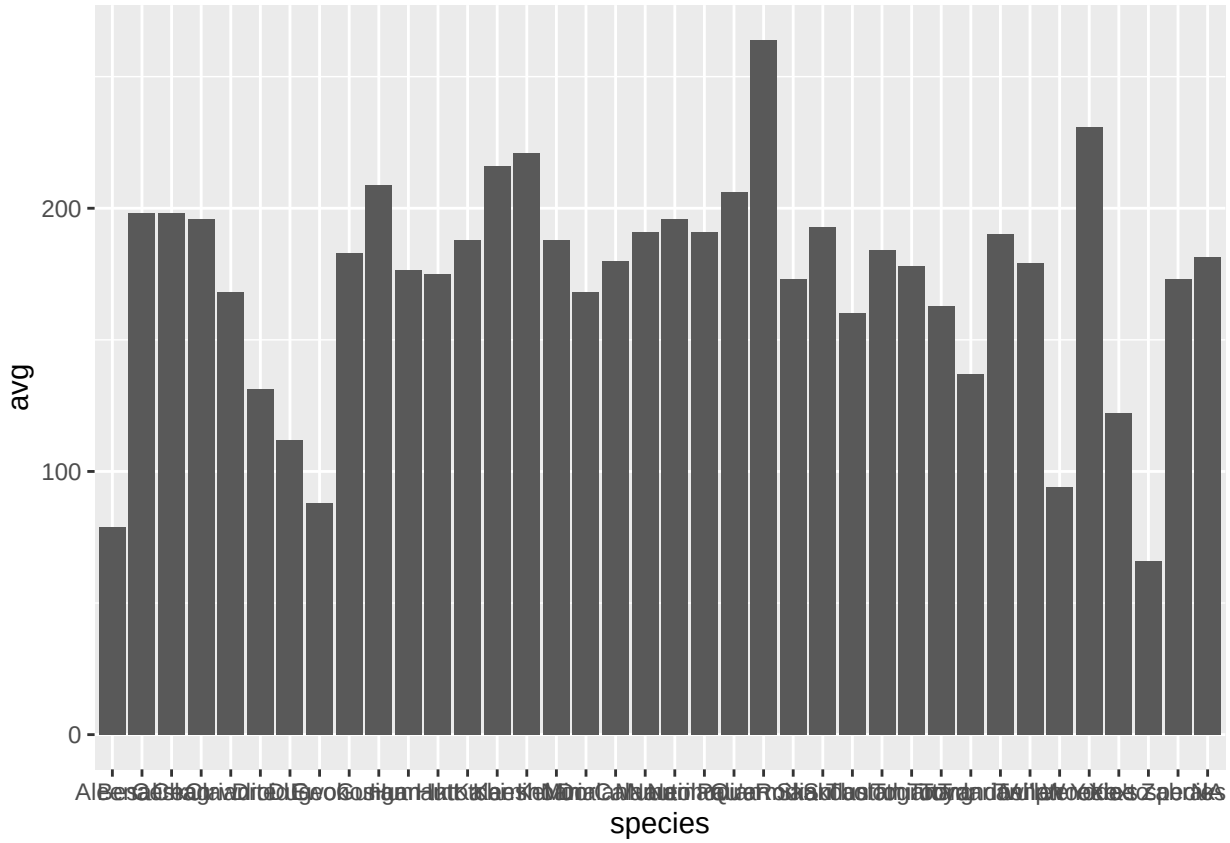
Scatterplot as columns

What is the average height for each species?

```
dat <- starwars %>%
  group_by(species) %>%
  summarise(avg = mean(height, na.rm = T))
```

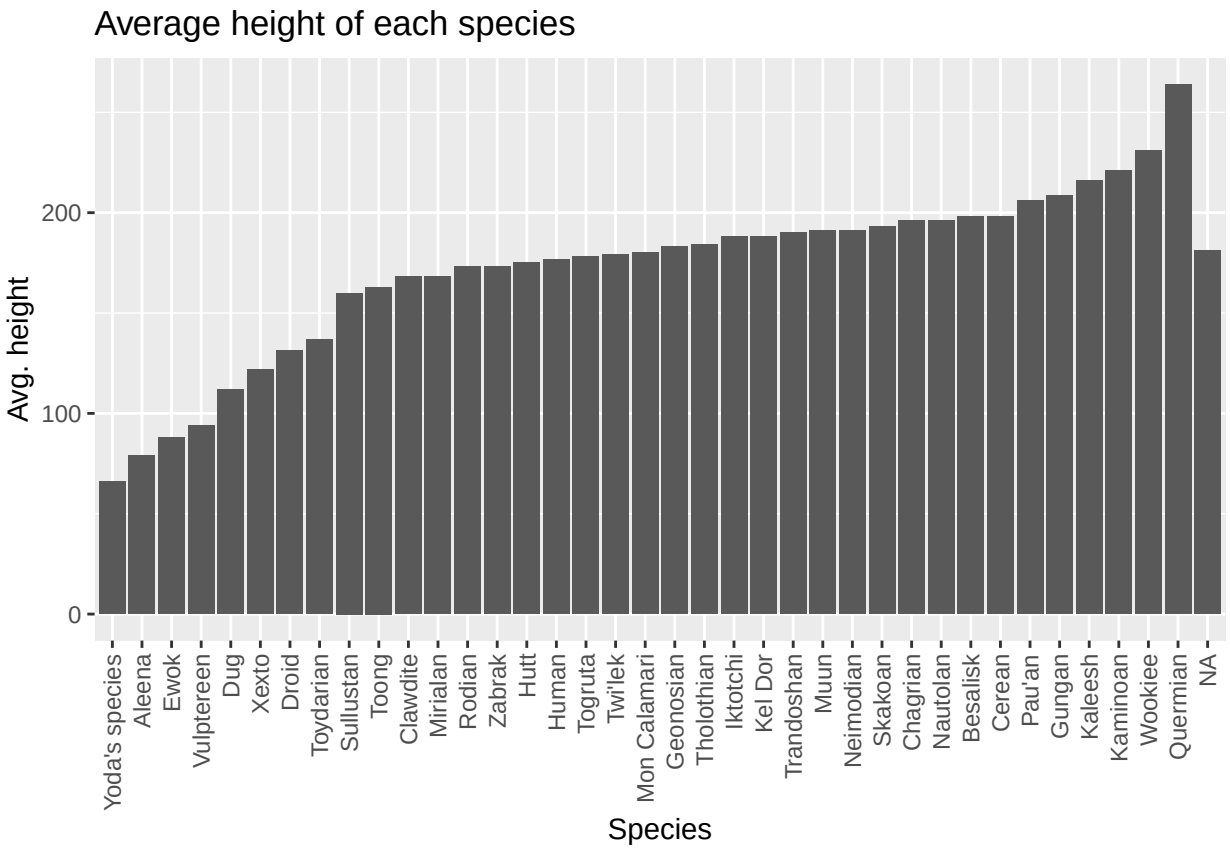
Let us visualize it

```
dat %>% ggplot(aes(x = species, y = avg)) +
  geom_col()
```



Let us sort the columns, add labels and rotate the x-axis labels:

```
dat %>%
  ggplot(aes(x = reorder(species, avg), y = avg)) +
  geom_col() +
  labs(title = "Average height of each species",
       x = "Species",
       y = "Avg. height"
    ) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

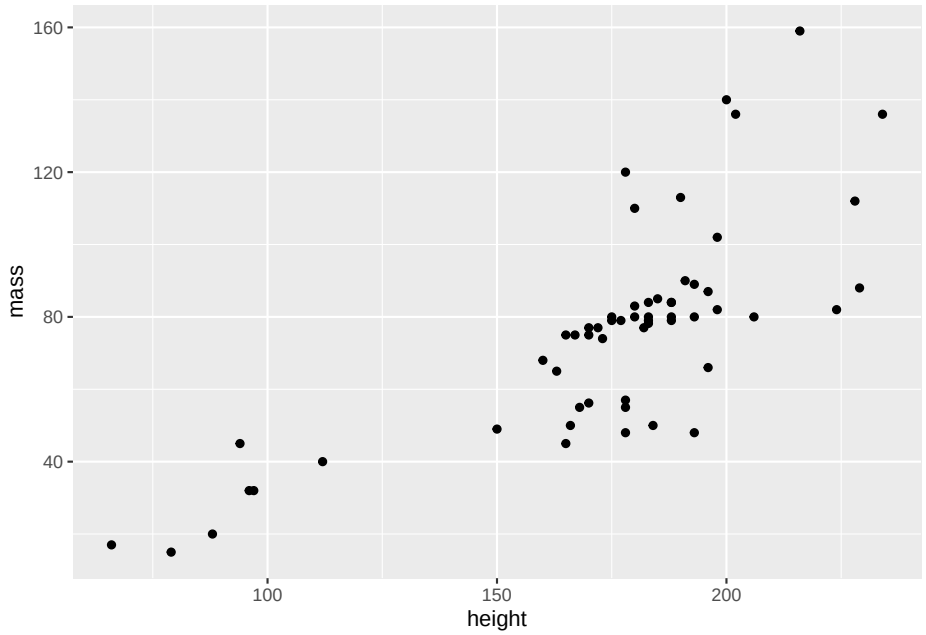


Visualizing relationships between numerical and numerical data

Scatterplot

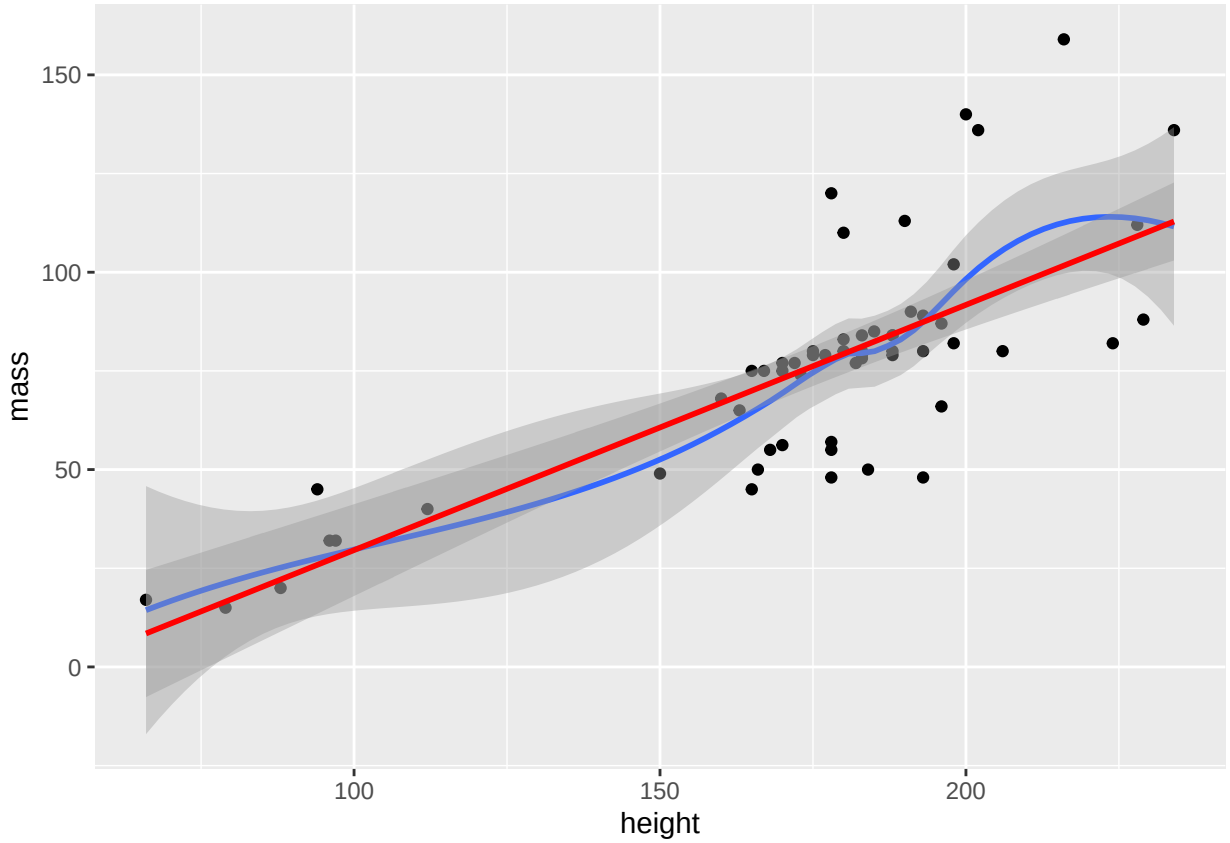
```
starwars_without_jabba <- starwars %>% filter(mass < 500)
```

```
ggplot(starwars_without_jabba, aes(x = height, y = mass)) +
  geom_point()
```



Let us try to add some lines

```
ggplot(starwars_without_jabba, aes(x = height, y = mass)) +
  geom_point() +
  geom_smooth() +
  geom_smooth(method = "lm", color = "red")
```


150
100
50
0
mass
100
150
200
height