

AE 04: Data visualization, Part 2

Visualizing Star Wars

Solutions

09.01.21

```
library(tidyverse)
library(viridis)

starwars <- read_csv("data/starwars.csv")

glimpse(starwars)

## Rows: 87
## Columns: 14
## $ name      <chr> "Luke Skywalker", "C-3PO", "R2-D2", "Darth Vader", "Leia Or~
## $ height    <dbl> 172, 167, 96, 202, 150, 178, 165, 97, 183, 182, 188, 180, 2~
## $ mass      <dbl> 77.0, 75.0, 32.0, 136.0, 49.0, 120.0, 75.0, 32.0, 84.0, 77.~
## $ hair_color <chr> "other", "other", "other", "none", "brown/auburn", "brown/a~
## $ skin_color <chr> "fair", "gold", "white, blue", "white", "light", "light", "~
## $ eye_color  <chr> "blue", "yellow", "other", "yellow", "brown", "blue", "blue~
## $ birth_year <dbl> 19.0, 112.0, 33.0, 41.9, 19.0, 52.0, 47.0, NA, 24.0, 57.0, ~
## $ sex        <chr> "male", "none", "none", "male", "female", "male", "female",~
## $ gender     <chr> "masculine", "masculine", "masculine", "masculine", "femini~
## $ homeworld  <chr> "Tatooine", "Tatooine", "Naboo", "Tatooine", "Alderaan", "T~
## $ species    <chr> "Human", "Droid", "Droid", "Human", "Human", "Human", "Huma~
## $ films      <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ vehicles   <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ starships  <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
```

We will continue using data about 87 characters in the *Star Wars* movie franchise. This analysis includes the following variables:

- **eye_color**: one of black, blue, brown, other, yellow
- **hair_color**: one of black, brown/auburn, none, other
- **height**: Height in centimeters (cm)

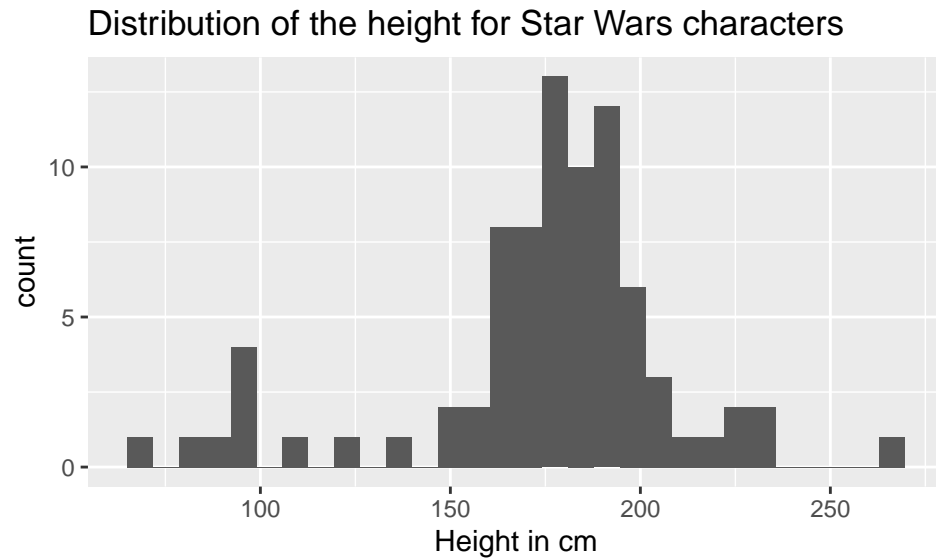
Step 1

Fill in the code below to create a histogram to visualize the distribution of **height**. **Once you have modified the code, remove the option `eval = FALSE` from the code chunk header.**

```
ggplot(data = starwars, mapping = aes(x = height)) +
  geom_histogram() +
  labs(x = "Height in cm",
       title = "Distribution of the height for Star Wars characters")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

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



- What is the shape of the distribution?

The shape of the distribution is unimodal and left-skewed.

Step 2

We can use the following code to calculate summary statistics for the distribution of height. We'll talk more about this code next week.

```
starwars %>%
  filter(!is.na(height)) %>%
  summarise(mean_height = mean(height), med_height = median(height),
            sd_height = sd(height), iqr_height = IQR(height))
```

```
## # A tibble: 1 x 4
##   mean_height med_height sd_height iqr_height
##   <dbl>      <dbl>    <dbl>    <dbl>
## 1      174.        180     34.8      24
```

- Which measure is best to describe the center of the distribution - mean or median? Briefly explain.

The median is the best measure to describe the center, because the distribution is left-skewed.

- Which measure is best to describe the spread of the distribution - standard deviation or IQR? Briefly explain.

The IQR is the best measure to describe the spread, because the distribution is left-skewed.

Step 3

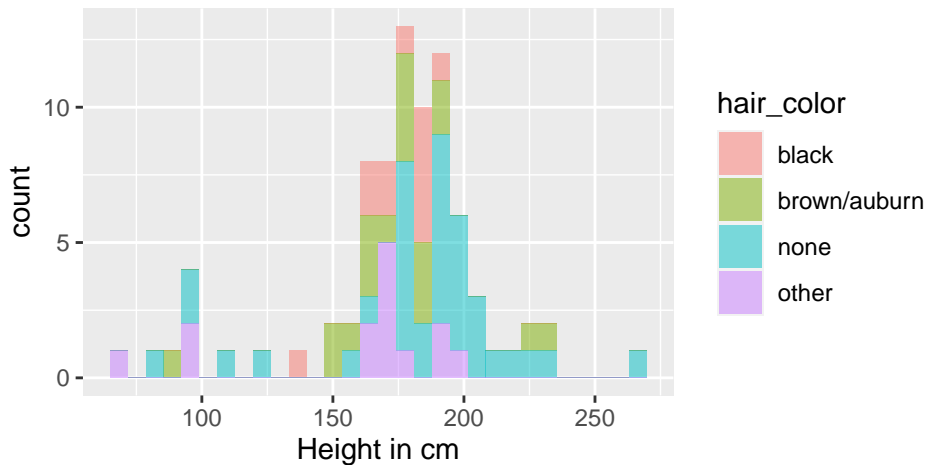
Do heights of characters differ by hair color? To answer this question, let's visualize the distribution of height for each category of hair color. Modify the code from Step 1 to fill in the color of the histogram based on hair color.

```
ggplot(data = starwars, mapping = aes(x = height, fill = hair_color)) +
  geom_histogram(alpha = 0.5) +
  labs(x = "Height in cm",
       title = "Distribution of the height for Star Wars characters",
       subtitle = "by Hair Color")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

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

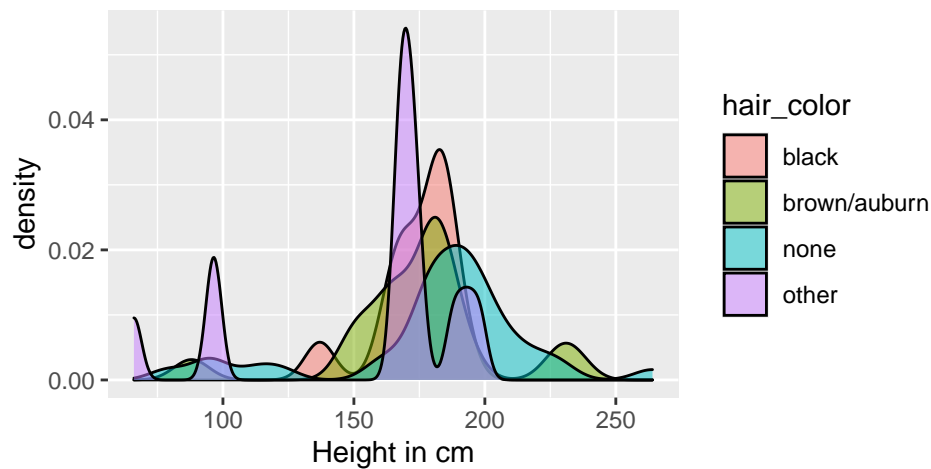
Distribution of the height for Star Wars characters
by Hair Color



```
ggplot(data = starwars, mapping = aes(x = height, fill = hair_color)) +  
  geom_density(alpha = 0.5) +  
  labs(x = "Height in cm",  
       title = "Distribution of the height for Star Wars characters",  
       subtitle = "by Hair Color")
```

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

Distribution of the height for Star Wars characters
by Hair Color



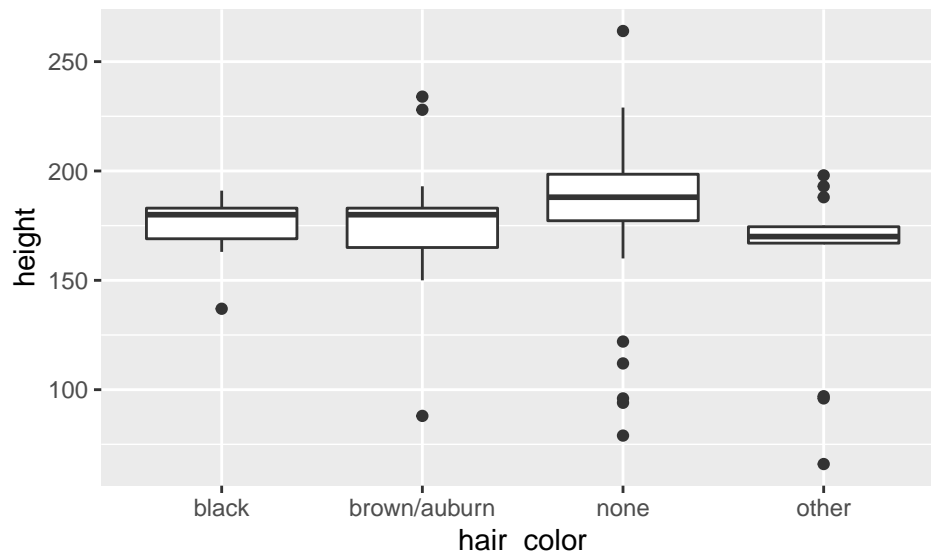
Step 4

Complete the code below to create side-by-side box plots to visualize the relationship between height and hair color. Once you have modified the code, remove the option `eval = FALSE` from the code chunk header.

```
# Add code here  
ggplot(data = starwars, mapping = aes(x = hair_color, y = height )) +
```

```
geom_boxplot()
```

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



Knit, commit, and push your changes to GitHub!

Resources

- ggplot2 reference page: https://ggplot2.tidyverse.org/reference/geom_histogram.html
- ggplot2 Cheat Sheet: <https://github.com/rstudio/cheat-sheets/raw/master/data-visualization.pdf>