AE 04: Data visualization, Part 2

Visualizing Star Wars

[INSERT YOUR NAME]

2021-08-31

```
library(tidyverse)
library(viridis)
```

We will continue using data about characters in the *Star Wars* movie franchise.

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

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 = ___, mapping = aes(x = ___)) +
  geom_histogram() +
  labs(title = "_____")
```

- What is the shape of the distribution?

Step 2

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

```
starwars %>%
  filter(!is.na(height)) %>% #remove observations with missing heights
  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?
- Which measure is best to describe the spread of the - standard deviation or IQR?

Step 3

Now let's consider the distribution of height for each category of hair color. Modify the code from Step 1 to create separate histograms with the color of each filled in based on the `hair_color`.

```
# add code here
```

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 = _____, y = _____ )) +
  geom_boxplot()
```

Step 5

- What feature(s) are apparent in both the histograms and side-by-side box plots?
- What feature(s) are apparent in the histograms that aren't apparent in the side-by-side box plots?
- What feature(s) are apparent in the side-by-side box plots that aren't apparent in the histograms?

Step 6

Finally, let's examine the relationship between hair and eye color. To do so, we'll use a segmented bar plot to visualize the distribution of eye color for each level of hair color. Fill in the code below to make the segmented bar plot. **Once you have modified the code, remove the option `eval = FALSE` from the code chunk header.**

```
ggplot(data = starwars, mapping = aes(x = _____, fill = ______)) +
  geom_bar(position = "fill") +
  labs(x = "_______",
       fill = "________",
       title = "_______",
       subtitle = "________") +
  scale_fill_viridis(discrete = TRUE) #apply viridis color palette
```

Note that we have used the `scale_fill_viridis` function from the **viridis** R package to apply the viridis color palette. This color palette makes the plots more accessible and more easily readable if printed in gray scale. Click here to read more about the viridis color palette.

Step 7

What are 2 observations about the relationship between hair and eye color based on the plot above?

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