ggplot2 Exercises

RICC International Young Investigator Training

11/16/2023

These exercises will give you some introductory experience with the `ggplot2`. Start by loading the `dplyr` and `ggplot2` library as well as the `murders` and `heights` data.

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
library(dslabs)
data(heights)
data(murders)
```

1. With `ggplot2` plots can be saved as objects. For example we can associate a dataset with a plot object like this

```r
p <- ggplot(data = murders)
```

Because `data` is the first argument we don't need to spell it out

```r
p <- ggplot(murders)
```

and we can also use the pipe:

```r
p <- murders %>% ggplot()
```

What is class of the object `p`?

2. Remember that to print an object you can use the command `print` or simply type the object. Print the object `p` defined in exercise one and describe what you see.

   a. Nothing happens.
   b. A blank slate plot.
   c. A scatterplot.
   d. A histogram.

3. Using the pipe `%>%`, create an object `p` but this time associated with the `heights` dataset instead of the `murders` dataset.

4. What is the class of the object `p` you have just created?

5. Now we are going to add a layer and the corresponding aesthetic mappings. For the murders data we plotted total murders versus population sizes. Explore the `murders` data frame to remind yourself what are the names for these two variables and select the correct answer. **Hint**: Look at `?murders`.

    a. `state` and `abb`.
    b. `total_murders` and `population_size`.
    c. `total` and `population`.
    d. `murders` and `size`.

6. To create the scatterplot we add a layer with `geom_point`. The aesthetic mappings require us to define the x-axis and y-axis variables, respectively. So the code looks like this:

```
murders %>% ggplot(aes(x = , y = )) +
  geom_point()
```

except we have to define the two variables `x` and `y`. Fill this out with the correct variable names.

7. Note that if we don't use argument names, we can obtain the same plot by making sure we enter the variable names in the right order like this:

```
murders %>% ggplot(aes(population, total)) +
  geom_point()
```

Remake the plot but now with total in the x-axis and population in the y-axis.

8. If instead of points we want to add text, we can use the `geom_text()` or `geom_label()` geometries. The following code

```
murders %>% ggplot(aes(population, total)) + geom_label()
```

will give us the error message: `Error: geom_label requires the following missing aesthetics: label`

Why is this?

    a. We need to map a character to each point through the label argument in aes.
    b. We need to let `geom_label` know what character to use in the plot.
    c. The `geom_label` geometry does not require x-axis and y-axis values.
    d. `geom_label` is not a ggplot2 command.

9. Rewrite the code above to use abbreviation as the label through `aes`

10. Change the color of the labels to blue. How will we do this?

    a. Adding a column called `blue` to `murders`.
    b. Because each label needs a different color we map the colors through `aes`.
    c. Use the `color` argument in `ggplot`.
    d. Because we want all colors to be blue, we do not need to map colors, just use the color argument in `geom_label`.

11. Rewrite the code above to make the labels blue.

12. Now suppose we want to use color to represent the different regions. In this case which of the following is most appropriate:

    a. Adding a column called `color` to `murders` with the color we want to use.
    b. Because each label needs a different color we map the colors through the color argument of `aes` .
    c. Use the `color` argument in `ggplot`.
    d. Because we want all colors to be blue, we do not need to map colors, just use the color argument in `geom_label`.

13. Rewrite the code above to make the labels' color be determined by the state's region.

14. Now we are going to change the x-axis to a log scale to account for the fact the distribution of population is skewed. Let's start by defining an object `p` holding the plot we have made up to now

```
p <- murders %>%
  ggplot(aes(population, total, label = abb, color = region)) +
  geom_label()
```

To change the y-axis to a log scale we learned about the `scale_x_log10()` function. Add this layer to the object `p` to change the scale and render the plot.

15. Repeat the previous exercise but now change both axes to be in the log scale.

16. Now edit the code above to add the title "Gun murder data" to the plot. Hint: use the `ggtitle` function.