Workshop 6: Intro ggplot & scatter plots

Joy Vaz

2022-11-29

1 Introduction

Hello!

For the first half of this workshop, we will be plotting datasets about HIV prevalence or incidence from countries around the world.

They number of persons that are estimated to be infected by HIV, including those without symptoms, those sick from AIDS and those healthy due to treatment of the HIV infection

Prevalence is a measure of the number of total cases in a year, and incidence is a measure of the number of new cases in a given time period.

The data also gives you the population for a given country-year.

The assignment should be submitted individually, but you are encouraged to brainstorm with partners.

The final due date for the assignment is Tuesday, November 29th at 23:59 PM UTC+2.

2 Get the assignment repo

To get started, you should download and look through the assignment folder.

1. First download the repo to your local computer here.

   You should ideally work on your local computer, but if you would rather work on RStudio Cloud, you can upload the zip file to RStudio Cloud through the Files pane. Consult one of the instructors for guidance on this.

2. Unzip/Extract the downloaded folder.

   If you are on macOS, you can simply double-click on a file to unzip it.

   If you are on Windows and are not sure how to "unzip" a file, see this image. You need to right-click on the file and then select "extract all".

3. Once done, click on the RStudio Project file in the unzipped folder to open the project in RStudio.

4. In RStudio, navigate to the Files tab and open the "rmd" folder. The instructions for your exercise are outlined there (these are the same instructions you see here).

5. Open the "data" folder and observe its components. Chose one of the two HIV datasets - would you like to analyze incidence or prevalence? Discuss with your partner and choose.

3 Load the data

Now that you understand the structure of the repo, you can load your chosen dataset.

In the code section below, **load in the needed packages**.

Now, **read the dataset into R**. Remember to use the `here()` function to allow your Rmd to use project-relative paths.

3.1 Step 1: Inspect the data types of your variables

The kinds of plots you can make is dependent on the data classes of your variable. Take a look at these types with `summary()` or `typeof()`.

Note that this dataset is structured very similarly to the `nigerm` dataset we used in Lesson 1.

3.2 Step 2: Choose your countries.

In this exercises you will be comparing patterns between countries. Since there are dozens of countries in the data, this may be too much to visualize at once. Choose 3-5 countries you would like to analyze.

(Hint: use `dplyr::filter()` to subset your data.)

You will be using `hiv_mini` to plot.

4 Creating a ggplot in layers

4.1 Time series plot

Using the subset you just made above, plot the number of cases over time as line graph. Line graphs are great for visualizing time series.

Think about what x and y aesthetic you need, and then which `geom_*()` function.

```
## [1] "ggplot(data = hiv_mini)"
```

Now, add color to this plot so that you can distinguish between countries. Then, increase the line width. Think about whether these aesthetic are mapping or fixed aesthetics!

Chaining together functions to add layers is a key feature of {ggplot2}. In the scatter plots lesson, we added `geom_smooth()` to `geom_point()` .

Think about what other types of plots besides line graphs can visualize the relationship between two continuous numeric variables.

You can check the *Cheatsheets* tab in your RStudio Help menu and browse the different geoms.

Next, add your another `geom_*()` layer to this plot.

Hint: there's a very common plot type that works for this!

Now, add at least 2 more aesthetics (fixed or mapped) to your plot.

The aesthetic we have covered so far are: - `color` - point color or point outline color

- `size` - point size
- `alpha` - point opacity
- `shape` - point shape
- `fill` - point fill color (only applies if the point has an outline)

Lastly, you can add a title to your plot with the `ggtitle()` function. See `?ggtitle` for more information.

4.2 Extra: explore scatter plots

Next you get to exercise your creativity! You task for this section is to use one of the datasets in the “extra” folder under “data”. Both these datasets contain continuous data that excellent for scatter plots.

1. **Create a scatter plot** with `geom_point()`, and show a relationship between two variables.
2. **Group or scale the points by a third variable**. You can use color, shape, are any other aesthetic to display information about an additional variable.
3. **Add a smoothing line!**

5 Submission: Upload HTML

Once you have finished the tasks above, you should **knit this Rmd into an HTML** and **upload it** on the assignment page.