INFO201 lab: Plotting

May 6, 2024

Instructions

This lab is *graded groupwork*. Please list the names of your group members on top of the lab! Note: only names that you list there will get the credit, so you may leave out those who do not contribute to the work.

In this lab we use data about COVID-19 cases in Scandinavia (Denmark, Finland, Norway and Sweden). The dataset *covid-scandinavia-wide* contains the following variables:

country
date
Confirmed cumulative number of confirmed COVID-19 cases by that date
Deaths cumulative number of COVID-19 deaths by that date

Good luck!

1 Covid over time

1. Load the dataset. Print a few lines. What does a row of data represent here?

2. What is the earliest and the most recent date in these data?

3. Next, your task is to plot the number of confirmed COVID cases in Denmark over time. (i.e. time on x-axis and confirmed cases on y-axis.)

   Explain what is a good way to display these data: scatterplot, line plot, barplot, or something else?

4. The number of confirmed cases reaches to millions. What do you think, what is a good way to represent such numbers?

5. Make the plot, ensure that the numbers are represented well in the way you suggested above.

6. Now make the same plot for all Scandinavian countries, denoting different countries by a different color.

---

2 Confirmed cases versus deaths

1. Make a scatterplot of confirmed cases versus deaths for Denmark.

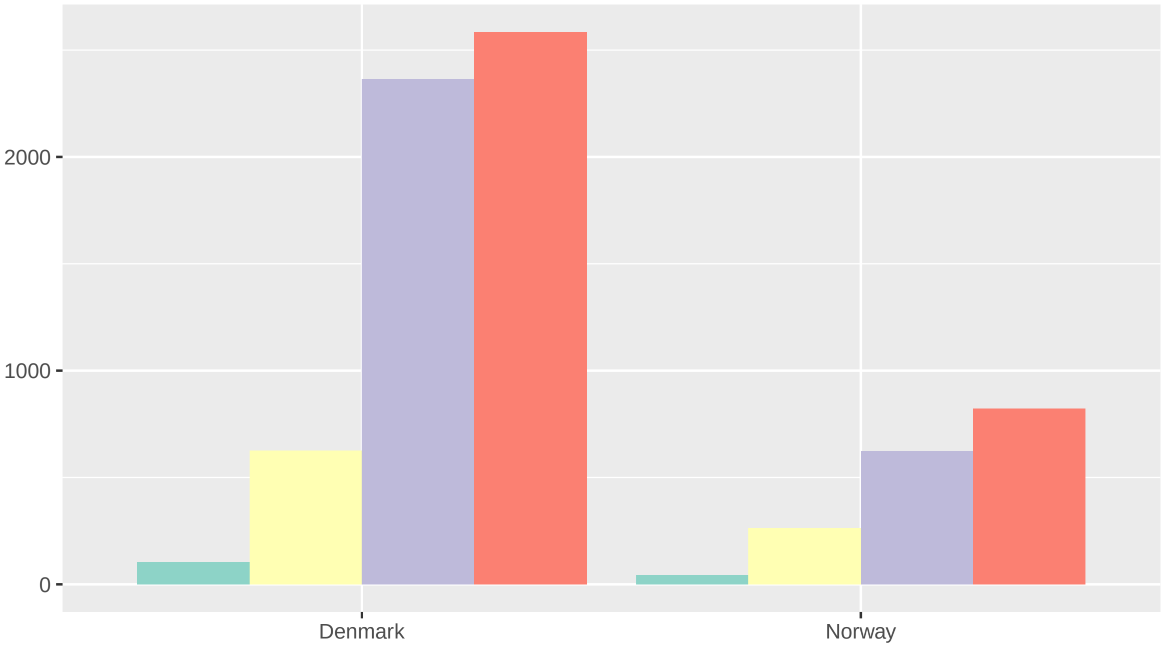
   Adjust the marker size and transparency to make it look good.

2. Now make a similar plot for all four countries, denoting different countries by different color.
3. Finally, replace the dots by paths (`geom_path()`) to show the trajectories.

---

3 Number of deaths over time

1. Extract the number of deaths for 2020-07-01, 2021-01-01, 2021-07-01 and 2022-01-01. Make a barplot of country versus deaths, where you mark the different dates with bars of different color. It should look something along these lines:



But you should include all countries and correct dates; what colors to pick is up to you.

2. Explain:
   (a) Is barplot a good way to display these figures? Why?
   (b) Do you want to use different colors for different countries or for different dates? Why do you think one option is better than the other?
   (c) Can you suggest a better plot instead of the barplot here?

---

4 Extra credit: Flight delays (1 EC pt = 0.1 credit pt)

In this section we use flights data, the same as in PS4.

1. Load data and ensure it looks good.

   Below we focus on flights to Seattle (dest = *SEA*) only. You can remove other flights.

2. Compute the percentage of flights that are delayed (at arrival) by more than 10 min for each carrier.

3. Explain how would you like to visualize these results. Line plot? Scatterplot? Barplot? Something else? Why?

4. Make the plot you suggested. Ensure it is appropriately labeled.

5. Now compute percentage of flights, delayed by more than 10 min, for each day of week through the whole 2013. Do this for both arrival and departure delays.

   Hint: function `wday()` in *lubridate* library can tell day of week of a date. Function `ISOdate(year, month, da` can make date out of year, month, and day.

6. Make a plot where you show both arrival delay and departure delay. Denote the weekdays by different color.

   Explain what do you think is a good way to make this plot.