Visualizing Data

Fundamentals of R Workshop - Homework 3

- All materials for the exercises below are available in the homework folder on Moodle.
- Please submit your answers as a HTML or a PDF file generated with Rmarkdown.
- Please make sure to include code chunks and plots in submitted the document.
- Please add titles, legends, and themes to your plots for clarity.

Ph.D. theses at the Graduate Institute III

The Institute has been the home for hundreds of PhD students over time. In this homework, we will visualize data to investigate how PhD theses at the Institute relate to gender and sustainability topics. The dataset contains the following variables:

Name	Description
`thesis_title`	Title of the Ph.D. thesis.
`thesis_ID`	ID of the Ph.D. thesis.
`thesis_year`	The year in which a thesis was submitted.
`thesis_author`	The author of the thesis.
`thesis_department`	Department in which the thesis was written.
`thesis_language`	Language in which the thesis was written.
`sus_degree`	Extent to which theses covers sustainability topics.
`gen_degree`	Extent to which theses covers gender-related topics.

Attention this is not the same dataset as previous weeks, please download the data from the correct Moodle folder

Question 1

a) Plot a histogram showing the distribution of PHD thesis over time (Tip: you could use base R here, see `?hist`). Why do you think we have so few thesis defenses in 2020?

b) Plot various histograms, side-by-side, showing the distribution of PHD theses over time by department (Tip: use `geom_histogram()` and facet the histograms by 'thesis_department').

Question 2

a) Do we have any NA values in the data? Please replace any NA values in the 'sus_degree' and 'gen_degree' variables with the value of 0 (Tip: you can use `mutate()` and `ifelse()`).

b) Please plot the extend to which theses covers sustainability ('sus_degree') and gender ('gen_degree') topics changed over time (Tip: smooth your lines to facilitate visualization, see `?geom_smooth`). Do the frequencies in which these topics appear in PhD theses increase over time?

c) Has the extent to which PhD theses cover sustainability ('sus_degree') and gender ('gen_degree') topics increased over time across all departments? Please plot the extend to which theses covers sustainability ('sus_degree') and gender ('gen_degree') topics changed over time by departments (Tip: remember `facet_wrap()`).

Question 3

a) Create a bar plot displaying the total number of theses written in each department. Please color the bars by the average sustainability score ('sus_degree') (Tip: you can use `group_by()` and `summarise()` to wrangle the data before plotting). Is this a good visualization?

b) Using the same bar plot as above (2a), please color the columns by the average gender score ('gen_degree') and reorder the columns in plot according to their total number (Tip: you can use `reorder()` for your x variable when plotting). How else could you improve this visualization?

Question 4

Plot all theses written before the year 2000, and after the year 2000, comparing the language they were written (Tip: use mutate to create separate categories for before and after 2000). Please interpret the plot.

Question 5

Create a scatter plot containing all theses that have a sustainability score ('sus_degree') bigger than 0. Please color the points by department ('thesis_department') and size them by their sustainability score ('sus_degree') (Tip: filter brfor plotting). What does the plot tell you?

Question 6

Please plot the relationship between thesis language and their gender score ('gen_degree'). Please interpret the plot.