HW 11: Data Visualization I (con't)

Graphical Analysis of Biological Data

By the end of this assignment, you should be able to achieve the following tasks in R:

- use R notebooks and R markdown;
- insert, write, and evaluate code chucks;
- use pipes;
- produce plots with `ggplot2`;
- visually analyze data for
  - outliers,
  - normality,
  - relationships among variables, and
  - interactions.
- use a typical workflow to wrangle and plot data;
- write custom functions, and
- confidently stage, commit, and push with Git.

These achievements belong to Learning Outcomes 2, 3, 4, 5, 6.

Click on any blue text to visit the external website.

Note: If you contact me for help or (better yet) open an issue in the public discussion forum, please include the code that is not working and also tell me what you have tried.

Preparation

- Open your `.Rproj` project file in RStudio.
- Create an `hw11` folder inside the same folder as your project file.
- Create a notebook file with the usual `<lastname>_hw11.Rmd` name and usual YAML header, and save it in your `hw11` folder.
- Right-click and save aegla_crabs.csv to your data folder.
- Add a code chunk to load the `tidyverse`, `here`, `ggally`, and `patchwork` libraries.
- Remember to format your code properly.
- Commit early. Commit often. Push regularly but at least push your completed assignment.
- Download and open A protocol for data exploration to avoid common statistical problems (free access) by Zuur et al. 2010.

```
library(tidyverse)
library(here)
library(GGally)
library(patchwork)
```

Habitat parameters for *Aegla* crabs

Data exploration: crabs

Aegla is a genus of South American freshwater crabs. Satterlee et al., as part of his research, collected water quality parameters. Those data are included in aegla_crabs.csv.

The variables are

- Site (sampling locations)
- Width, Depth (stream width and depth)
- Flow (rate of water movement)
- AT, WT (air and water temperature)
- pH (pH; What did you expect me to say, log molar hydrogen ion concentration?)
- TDS (total dissolved solids)
- Cond (conductivity)
- N, Ni, Nt (Nitrogen, Nitrate, Nitrite concentrations)
- Phos, Mg, Ca (phosphate, magnesium, calcium concentrations)
- Crab, Shrimp, Aeglam, AeglaFe, AeglaFo (number of shrimp and crabs sampled)

Your goal is simple but it will take time.

- Import the raw data from `aegla_crabs.csv`.
- Use `select()` to remove the Site, AT, WT, Crab, Shrimp, Aeglam, AeglaFe, AeglaFo columns.
- Use `drop_na()` to remove the one row with missing water quality data.
- Create a `row_order` dummy variable like we did for the sparrows.
- You must use the pipe (`%>%`) to write efficient code.
- **Explore the data.** There are three apparent outliers among the variables but only one that I think is an actual outlier. Find them and justify your choices.
- Use `ggpairs` to explore the relationships among all the variables. `ggpairs` is a great place to start looking for outliers but it's a start, not a destination. You must do more.
- **Describe the results** that you see in your figures. Use the notebook to describe your interpretations of the data. Your descriptions must be more than single sentences although they do not have to be lengthy paragraphs. Two or three sentences is probably sufficient for most cases. I am looking for evidence that you studied and thought about the data. *I want you to think like the scientist you are becoming.*
- You do not need to describe every panel produced from `ggpairs`. Use it as a guide to help you understand the data.
- Choose four variables to make Cleveland plots, and make a 2x2 grid with the four plots, using `patchwork`.
- Choose three different variables to make histograms. Play with the `bins` or `binwidths` argument that you think reveals the data well. *Hint:* the default `bins = 30` that `geom_histogram` is usually not a suitable choice. Save each histogram to a unique variable.
- Use the same three variables to produce density plots. Save each plot to a unique variable.
- Use 'patchwork to produce a 2 column matrix with the histograms in the left column and the corresponding density plots in the right column.

The *Aegla* crabs data is not suitable for testing for interactions. I'll see if I can find an opportunity to do that in a future assignment. I know you are thrilled.

---

et Vóila