Exploratory Data Analysis in R with Tidyverse

Session 6: Case Study

instats

Revisiting the EDA Process

- EDA is a structured way to explore and understand your data

- In this case study, we will follow three major stages:
  - **Load and clean**: standardize column names, check for missing values
  - **Understand structure**: explore variable types, counts, and summaries
  - **Uncover patterns**: relationships between traits, grouped summaries, and visualizations

Core EDA Tools in Action

- Use `janitor`, `dplyr`, `tidyr`, and `purrr` for transformation and summaries

- Use `ggplot2` and `plotly` to visualize structure, relationships, and distributions

- Revisit key functions (and many more!):
  - `summarize()`, `mutate()`, `pivot_longer()`
  - `group_by()`, `count()`, `map_df()`
  - `ggplot()`, `plot_ly()`, `ggplotly()`

Thinking with Data: From Traits to Trends

- Categorical traits → use bar charts and counts
- Numeric traits → use histograms and boxplots
- Combined views → use grouped or faceted plots
- Always return to EDA's key questions:
  - What's typical?
  - What's unusual?
  - What varies across groups?