# Exploratory Data Analysis in R with Tidyverse

## Session 2: Data Cleaning Fundamentals with dplyr, tidyr, and janitor

instats

# Selecting and Filtering for Relevant Observations

- Use `select()` to focus on relevant variables for your analysis
- Use `filter()` to narrow the dataset to just the rows you need
- Combine with `arrange()` and `slice_sample()` to inspect sorted or random subsets
- Enables fast inspection and validation of data quality

# Creating New Variables for Analysis

- Use `mutate()` to derive new insights from existing columns
- Create binary flags like `high_energy_dance`
- Use `case_when()` for multiple conditional groupings like `popularity_group`
- Essential for feature engineering and summarization

# Reshaping Data for Deeper Analysis

- Use `pivot_longer()` to stack multiple columns into one (long format)
- Use `pivot_wider()` to spread grouped counts across columns (wide format)
- Enables advanced visualization and comparison across features or categories

# Separating and Uniting Text Columns

- Use `separate()` to split one column into many (e.g., artists list)
- Use `unite()` to combine multiple columns into one (e.g., “track by artist”)
- Great for cleaning and structuring messy text fields

# Tidying and Recoding Variables

- Use `janitor::clean_names()` to standardize messy column names
- Use `remove_empty()` to remove empty columns
- Use `mutate()` and `case_match()` to recode or reorder categorical values
- Critical for making your data analysis-ready