Exploratory Data Analysis in R with Tidyverse

Session 3: Managing Dates, Strings, and Categories with lubridate, stringr, and forcats

instats

Extracting Meaning from Strings

- Use `str_detect()` to search for keywords like “love” in track names
- Use `str_length()` and `str_count()` to quantify text patterns
- Convert strings to lowercase for consistent matching
- Summarize counts across groups with `group_by()` and `summarize()`

Cleaning Up Track Titles with Regex

- Use `str_remove_all()` to strip out parentheses and featured credits
- Use regular expressions like `"\\(.*?\\)"` for non-greedy matching
- Extract phrases with `str_extract()` and clean them with `str_replace_all()`
- Get cleaner track names for visualizations and summaries

Working with Artists and Categorical Variables

- Use `separate_rows()` to split multi-artist entries into long format
- Count artist appearances and identify those crossing genres
- Use `fct_lump()` to group rare categories into “Other”
- Use `fct_relevel()` to control the order of factor levels

Parsing Dates and Grouping by Time

- Use `make_date()` to construct full dates from parts
- Use `floor_date()` to create time bins (e.g. every 6 hours)
- Extract month or weekday labels with `month()` and `wday()`
- Group and summarize by date-time intervals

Visualizing Categorical Preferences and Trends

- Use `count()` with `group_by()` to tabulate responses by category
- Reorder with `fct_infreq()` to sort by most common responses
- Collapse rare entries with `fct_lump()`
- Use `geom_col(position = "dodge")` to compare group counts