Exploratory Data Analysis in R with Tidyverse

Session 1: Foundations of EDA in R with the tidyverse

instats

What is EDA?

- Process of exploring data to uncover patterns and insights
- Essential for understanding data before modeling
- Supports better questions, cleaner data, and stronger decisions

Importance of EDA

- Early Error Detection
  - Identifies data anomalies
  - Corrects erroneous assumptions
- Enhancing Data Understanding
  - Summarizes main characteristics
  - Reveals underlying structure

Installing and Loading Key Packages

- Core tidyverse packages: dplyr, ggplot2, tidyr, etc.
- Additional tools: janitor, plotly, moderndive, purrr
- Load data using `data()`

Meet the Data: `spotify_by_genre`

- A Rich Dataset for EDA Practice
- 6,000 Spotify tracks across six genres
- Mix of metadata, audio features, and popularity indicators
- Designed for music trend analysis and modeling

Core EDA Skills: Viewing, Cleaning, and Sampling

- Key Actions Before Analysis
- `glimpse()` for quick structure review
- `distinct()` to remove duplicates
- `count()` and `slice_sample()` for basic exploration

Wrapping Up: Variable Types and Summary Tools

- Summary Stats and Quick Visuals
- Use `class()` and `map()` to inspect variable types
- `summarize()` or `tidy_summary()` for quick stats
- `ggplot()` with `geom_histogram()` to visualize distributions