Welcome to

instats

The Session Will Begin Shortly

START

Statistics in R with Tidyverse

Session 2: Data Visualization using ggplot2

instats

Introduction to Data Visualization

- Insights that raw data alone cannot provide
- `ggplot2` package based on Grammar of Graphics by Leland Wilkinson
- Visualizations help to identify outliers, distributions, and relationships

Grammar of Graphics

- A statistical graphic maps data variables to aesthetic attributes
- Key components:

1. `data`: The dataset
2. `geom`: The geometric objects (points, lines, bars)
3. `aes`: Aesthetic attributes like position, color, shape, size

- Create visualizations by layering these components in `ggplot()`

The Five Named Graphs

- Essential tools for data visualization
- Scatterplots, linegraphs, histograms, boxplots, and barplots
    - Each type works best for different data relationships and distributions
    - Goal is to uncover trends, patterns, and outliers in data

Scatterplots

- Display relationships between two numerical variables
- Using `geom_point()`
- Customizing points (`color`, `shape`, `size`)
- **Tip**: Handling overplotting
    - `alpha` transparency
    - jittering with `geom_jitter()`

Linegraphs

- Display trends over time or relationships between two sequential variables
- Use `geom_line()`
- Commonly used for time-based data (hours, days, weeks, etc.)
- **Tip**: Avoid using linegraphs when the x-axis variable has no inherent order

Histograms

- Display the distribution of a single numerical variable
- Use `geom_histogram()`
- Visualize data spread, center, and frequency of values
- **Tip**: Adjust bin width or number of bins for better data representation

Boxplots

- Summarize numerical data using quartiles and medians
- Use `geom_boxplot()`
- Effective for identifying data spread and detecting outliers
- **Tip**: Use boxplots for comparing distributions across groups

Barplots

- Display the distribution of a categorical variable's frequencies
- Use `geom_bar()`or `geom_col()`
- Barplots are ideal for comparing frequencies of categories or groups
- Tip: Use `geom_bar()` for raw (uncounted) data and `geom_col()` for pre-counted data

Demo & Exercises

Q & A

STOP