Introduction to Statistics in R

Workshop 2 – September 23, 2025

Statistical tests and models in R

Hands-on tutorials to help build skills required in Precision Child Health, data science and bioinformatics for SickKids staff and trainees.

Topics include:

- Data visualization
- Correlations
- T-tests
- Linear regression

No prior knowledge of programming is required. We encourage you to bring your laptop to practice data analysis steps under our guidance.

To register please visit:
http://ccm20250923.eventbrite.com/

Hands-on Data Science Tutorials for Biologists

Tuesday Sep 23, 2025
12:00 – 1:00 pm

Multi-media Room
PGCRL 3rd floor
or online (hybrid event)

Supported by :
Centre for Computational Medicine (ccm.sickkids.ca)
Digital Research Alliance of Canada (alliancecan.ca)
Compute Ontario (computeontario.ca)

CCM Overview

Note that prior knowledge of R is recommended to get the most out of this workshop series.

Workshop 1	Data exploration and visualization
Workshop 2	Statistical tests and models in R

Check out other workshop series hosted by CCM at https://ccm.sickkids.ca/bioinformatics-training/

Table of contents

01

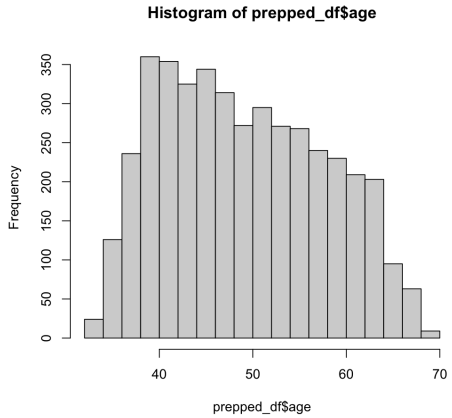
Data visualization

Base R vs ggplot2

Base R

Useful in a pinch but doesn't generate the prettiest figures.

`plot()` and `hist()` are useful for taking a quick look at your data.



ggplot2

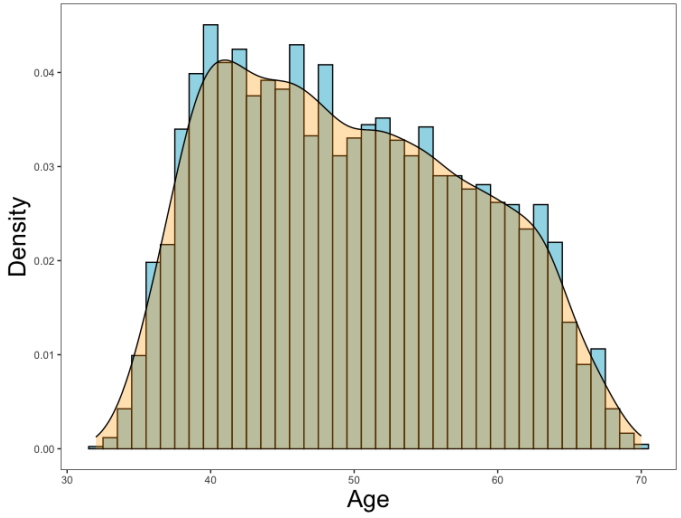
ggplot2 is a package based on the grammar of graphics.

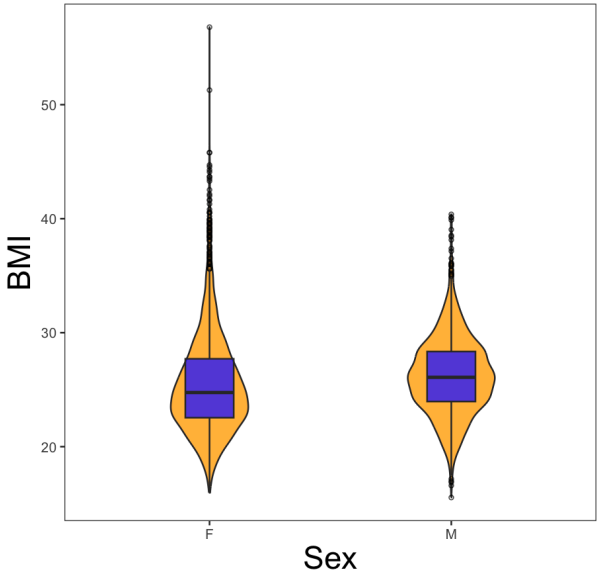
You can create your figure with some basic components:

1. **Data**
2. **Mapping**
3. **Layers**
4. Scales
5. Facets
6. Coordinates
7. Theme

Distributions

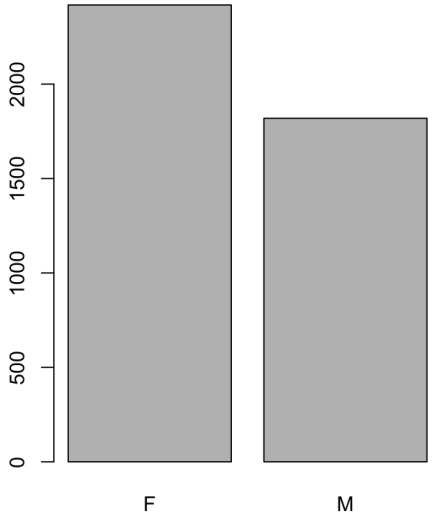
Look at the distribution of data using histograms, density plots, boxplots, and violin plots (among others).



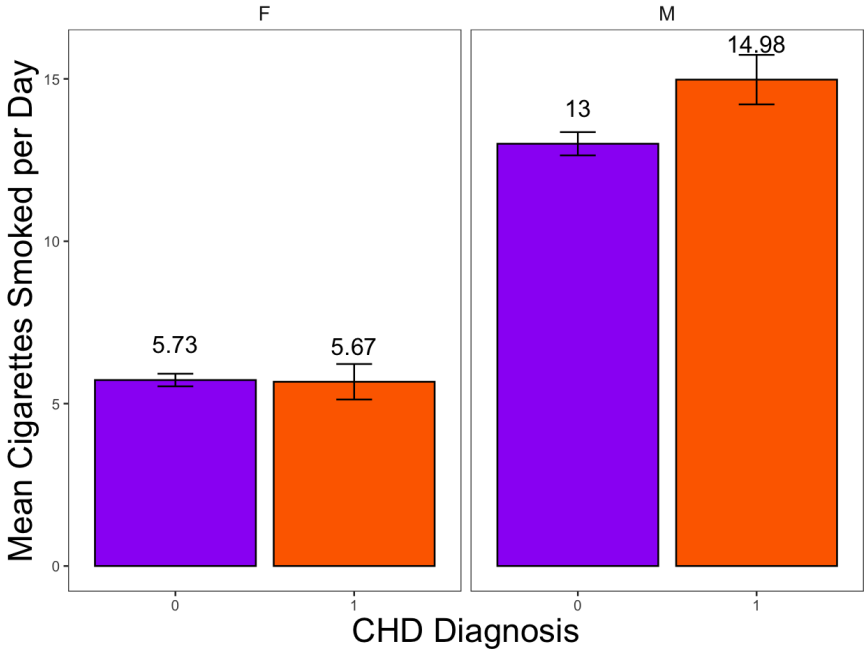


Bar Charts

Compare groups using bar charts.



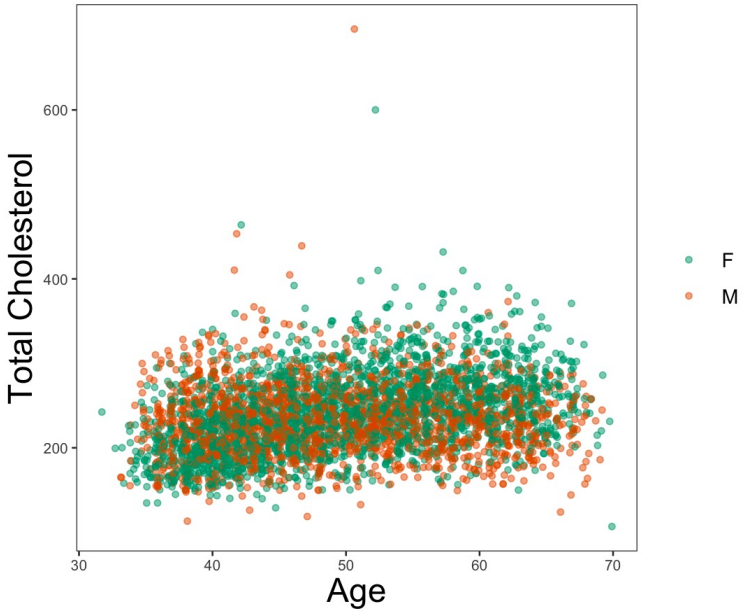
Base R

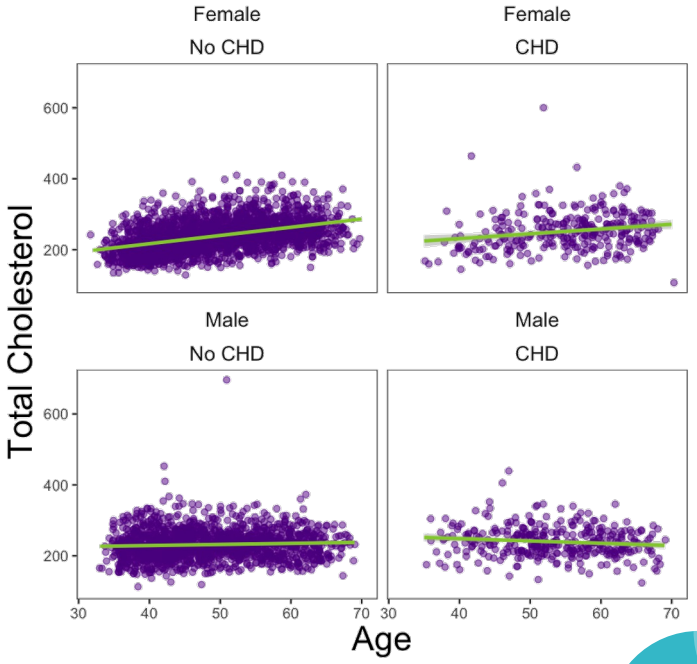


ggplot2

Scatterplots

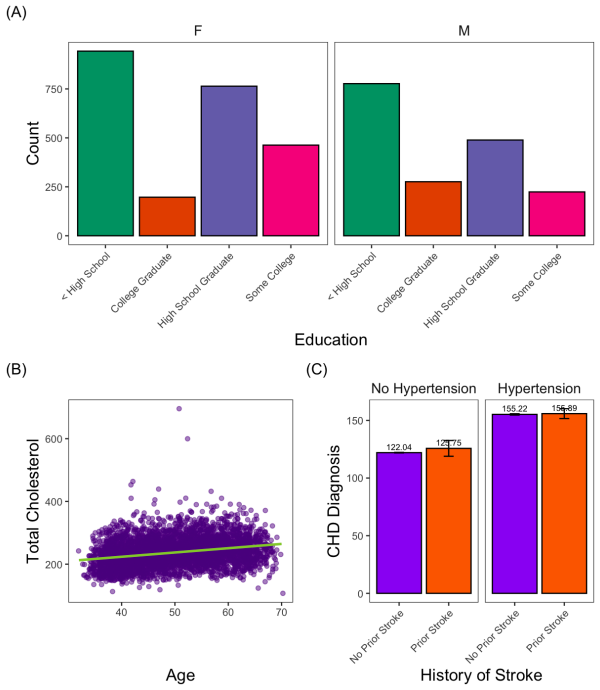
Plot the relationship between two continuous variables using scatterplots.





Panel figures

You might have multiple plots that you want to put into a single figure with panels. You can do this using the `patchwork()` package.

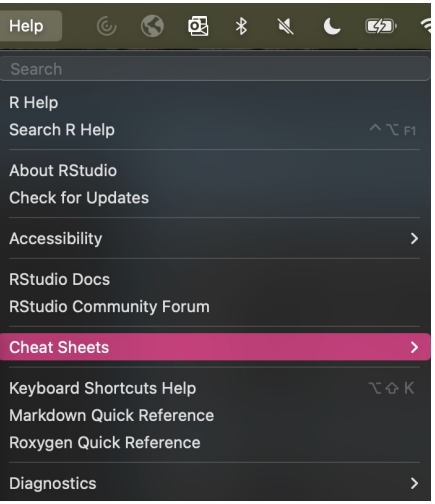


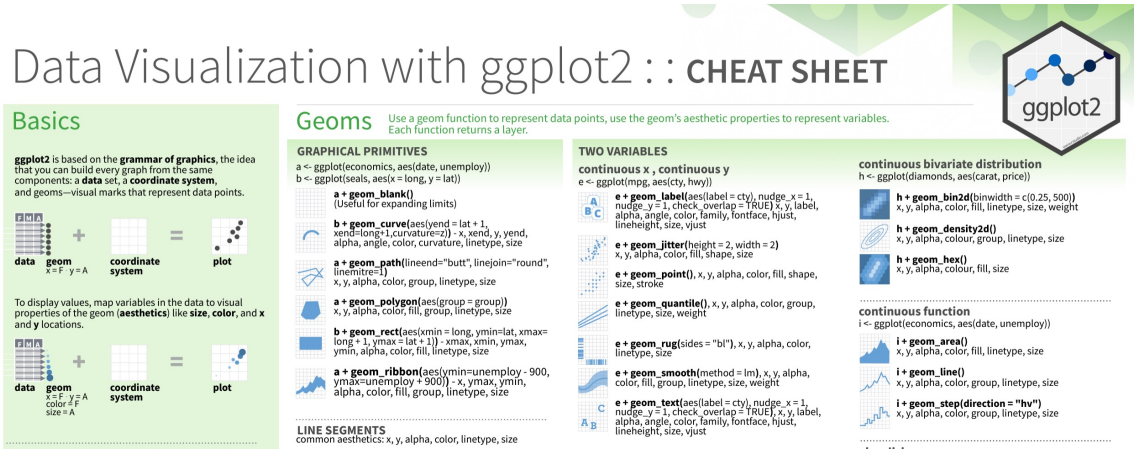
Resources

ggplot2 cheat sheet:
https://www.maths.usyd.edu.au/u/UG/SM/STAT3022/r/current/Misc/data-visualization-2.1.pdf

color palette reference:
https://sape.inf.usi.ch/quick-reference/ggplot2/colour

R Studio cheat sheets: Help > Cheat Sheets > …




Data Visualization with ggplot2 : : CHEAT SHEET

ggplot2

Basics

ggplot2 is based on the grammar of graphics, the idea that you can build every graph from the same components: a data set, a coordinate system, and geoms—visual marks that represent data points.

data geom coordinate system plot

To display values, map variables in the data to visual properties of the geom (aesthetics) like size, color, and x and y locations.

data geom coordinate system plot

Geoms Use a geom function to represent data points, use the geom's aesthetic properties to represent variables. Each function returns a layer.

GRAPHICAL PRIMITIVES

a <- ggplot(economics, aes(date, unemploy))
b <- ggplot(seals, aes(x = long, y = lat))

a + geom_blank() (Useful for expanding limits)

b + geom_curve(aes(yend = lat + 1, xend=long+1,curvature=z)) - x, xend, y, yend, alpha, angle, color, curvature, linetype, size

a + geom_path(lineend="butt", linejoin="round", linemitre=1) x, y, alpha, color, group, linetype, size

a + geom_polygon(aes(group = group)) x, y, alpha, color, fill, group, linetype, size

b + geom_rect(aes(xmin = long, ymin=lat, xmax= long + 1, ymax = lat + 1)) - xmax, xmin, ymax, ymin, alpha, color, fill, linetype, size

a + geom_ribbon(aes(ymin=unemploy - 900, ymax=unemploy + 900)) - x, ymax, ymin, alpha, color, fill, group, linetype, size

LINE SEGMENTS
common aesthetics: x, y, alpha, color, linetype, size

TWO VARIABLES

continuous x , continuous y

e <- ggplot(mpg, aes(cty, hwy))

e + geom_label(aes(label = cty), nudge_x = 1, nudge_y = 1, check_overlap = TRUE) x, y, label, alpha, angle, color, family, fontface, hjust, lineheight, size, vjust

e + geom_jitter(height = 2, width = 2) x, y, alpha, color, fill, shape, size

e + geom_point(), x, y, alpha, color, fill, shape, size, stroke

e + geom_quantile(), x, y, alpha, color, group, linetype, size, weight

e + geom_rug(sides = "bl"), x, y, alpha, color, linetype, size

e + geom_smooth(method = lm), x, y, alpha, color, fill, group, linetype, size, weight

e + geom_text(aes(label = cty), nudge_x = 1, nudge_y = 1, check_overlap = TRUE), x, y, label, alpha, angle, color, family, fontface, hjust, lineheight, size, vjust

continuous bivariate distribution

h <- ggplot(diamonds, aes(carat, price))

h + geom_bin2d(binwidth = c(0.25, 500)) x, y, alpha, color, fill, linetype, size, weight

h + geom_density2d() x, y, alpha, colour, group, linetype, size

h + geom_hex() x, y, alpha, colour, fill, size

continuous function

i <- ggplot(economics, aes(date, unemploy))

i + geom_area() x, y, alpha, color, fill, linetype, size

i + geom_line() x, y, alpha, color, group, linetype, size

i + geom_step(direction = "hv") x, y, alpha, color, group, linetype, size