Introduction to Statistics in R

Workshop 1 – September 16, 2025

CCM

Data exploration and visualization 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 exploration in R
- Preparing data for analysis
- Data visualization using ggplot

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://ccm20250916.eventbrite.com/

Hands-on Data Science Tutorials for Biologists

Tuesday Sep 16, 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 exploration and cleaning

Framingham Heart Study

One of the longest prospective epidemiological studies of cardiovascular disease and its risk factors.

Began in 1948 in Framingham, MA.

- Initially enrolled 5209 men and women aged 29–62 years old.
- Followed them over time, with assessments every 2 years.
- Data collection for the original cohort ended in 2014.
- Over time, the researchers incorporated offspring and their spouses into the study.
- Today, they have data from three generations of participants as well as more ethnically diverse cohorts.

Framingham Heart Study

Examinations included:

- Detailed medical history
- Physical exams
- Lab tests
- Lifestyle and habits
- Noninvasive imaging

Our dataset: subset of the FHS from Kaggle

- 4000+ records and 16 variables.

(https://www.kaggle.com/datasets/captainozlem/framingham-chd-preprocessed-data)

Variable	Description	Class/type
male	Sex (male or female)	Binary
age	Age	Continuous
education	Education (0-11 years, HS or GED, some uni, uni grad+)	Ordinal/categorical
currentSmoker	Whether the patient is currently a smoker	Binary
cigsPerDay	Number of cigarettes smoked per day, on average	Continuous
BPMeds	Whether the patient is on blood pressure medication	Binary
prevalentStroke	Whether the patient previously had a stroke	Binary
prevalentHyp	Whether the patient is hypertensive	Binary
diabetes	Whether the patient has diabetes	Binary
totChol	Total cholesterol level	Continuous
sysBP	Systolic blood pressure	Continuous
diaBP	Diastolic blood pressure	Continuous
BMI	Body mass index	Continuous
heartRate	Heart rate	Continuous
glucose	Glucose level	Continuous
TenYearCHD	10-year risk of coronary heart disease	Binary

Data exploration

Some common and/or important data cleaning steps include making sure that:

1. All variables are in the right format and/or correctly classified.
2. Duplicate or missing data are dealt with.
3. Variables of interest have been transformed appropriately.
4. Outliers are handled.
5. Data are validated.

Data exploration

1. Look at the data and its structure – are all the variables in the right format?

`str()`

2. Tabulate data

`table(df$column)` or `table(df$column1, df$column2, ...)`

3. Modify variables

```
mutate(column = function(column))
mutate(df, across(columns, function))
```

4. Transform variables – center, scale, log-transform

```
log(x, base = exp(1))
scale(x, center = TRUE, scale = TRUE)
```

Data exploration

5. Subset dataframes

`filter()` rows according to some condition

`select()` columns to keep

```
subset(df, subset = row_vals == “abc”, cols = X)
```

6. Combine vectors and dataframes

`cbind()` to join by columns

`rbind()` to join by rows

`merge()` to join dataframes

7. Reshape dataframes

`pivot_wider()` each row is a participant/patient

`pivot_longer()` each row is an observation, with multiple rows per participant

Data exploration

8. Handy functions, pipes, and operators

`ifelse()` for conditional element selection
`case_when()` as a vectorized version of multiple `ifelse()` statements, uses formula notation (more on that later)

`%>%` passes the result of the function on the left as input to the next function
`%in%` to compare vectors

Create a summary table

Use `summarize()` to create a dataframe with summaries of desired variables. For example, what if you're only interested in the educational attainment and smoking habits of people who go on to develop heart disease?

	TenYearCHD	education_class	n_participants	age	currentSmoker	prop_smokers	cigsPerDay
1	0	< High School	1397	51	645	0.46	8.39
2	0	College Graduate	403	47	204	0.51	9.09
3	0	High School Graduate	1106	47	598	0.54	9.67
4	0	Some College	599	48	275	0.46	7.70
5	1	< High School	323	56	160	0.50	9.93
6	1	College Graduate	70	53	36	0.51	11.97
7	1	High School Graduate	147	52	81	0.55	11.14
8	1	Some College	88	53	46	0.52	10.74