1 Programming

2 Data Organization

A central framework of data organization in R is the **tidyverse**, especially the **package dplyr**. Alongside it, **tidyr** focuses on reshaping data between wide and long formats, handling missing values, and ensuring that datasets follow the "tidy data" structure where each variable is a column and each observation is a row. For data import and string handling, packages like **readr** and **stringr** are commonly used, making it easier to read large datasets and clean textual variables in a consistent way. The design philosophy of the tidyverse is to make data preparation readable, chainable, and close to natural language through the pipe operator %>%. For large-scale data processing, packages like **data.table** are often preferred because they are faster and more memory-efficient, especially for big datasets. These are functions (**Rcode**) I often use:

- **filter()** means selecting a subset of observations based on conditions, such as keeping only patients older than 60 or only records from a specific time period.
- **select()** means choosing specific variables from a dataset, such as keeping only age, sex, and outcome while removing unnecessary columns.
- **group_by()** divides data into subsets based on one or more categorical variables, such as grouping by gender.
- **summary** means aggregating data within groups or across the entire dataset to compute statistics such as mean.
- **mutate()** is used to create new variables or modify existing ones based on transformations of other columns. For example, you can compute a new risk score, log-transform a variable, or combine existing variables into a new feature.
- **drop_na()** is used to remove rows that contain missing values (NA).
- **is.na()** is used to detect missing values. It returns a logical vector indicating whether each element is NA, and is often used for filtering, counting missing values, or creating indicators of missingness.
- **which(!is.na(x))** is used to identify the positions of all elements in x that are not *NA*.
- **apply** is a function used to perform an operation over the rows or columns of a matrix or array without writing explicit loops.

3 Data Visualization

In R, there are several ways to create visualizations (e.g., ggblend, ggnewscale, gganimate), and ggplot2 is one of the most widely used and influential approaches (**R Graph Gallery**).

- For **distributional analysis**: histograms, density plots, frequency polygons, boxplots, violin plots, ridgeline plots, dot plots, and empirical cumulative distribution function (ECDF) plots, all of which help reveal the shape, spread, and skewness of data.
- For textbfrelationships between variables: scatter plot, bubble charts, regression lines, smoothing curves, confidence bands, two-dimensional density plots, contour plots, and hexbin plots.
- For textbfcomparisons across categories: bar charts (count and summarized), grouped and stacked bar charts, proportional (100%) bar charts, lollipop charts, Cleveland dot plots, and mosaic plots.

- For **time series data**: line plots, area charts, stacked area charts, streamgraphs, and ribbon plots.
- For **multivariate visualization**: heatmaps, tile plots, scatterplot matrices, parallel coordinate plots, and faceted plots.
- For **spatial data**: choropleth maps, point maps, path maps, polygon maps, and raster maps.
- For **statistical summaries and uncertainty visualization**: error bar plots, pointrange plots, linerange plots, and confidence/credible interval visualizations.
- Sankey diagrams and Alluvial plots visualize flows between categories, where link thickness represents magnitude or proportion.

ggplot2 supports advanced visualizations such as QQ plots, PP plots, slope graphs, waterfall charts, funnel plots, dendrograms, and network visualizations. To create customized graphs, you can adjust **colors**, **plot styles**, **combining plots**, and textbftheme settings.

4 Modeling

In modern data analysis and machine learning, two fundamental requirements are generalization performance and reproducibility (**Ref**).

- **Generalization performance** refers to how well a model trained on observed data performs on unseen data, capturing the ability to extract true underlying patterns rather than overfitting noise. Cross-validation is commonly used to estimate this.
- **Reproducibility** refers to whether results can be reproduced using the same data, code, and procedures. It ensures scientific reliability and transparency.

5 Generalized Linear Model (GLM)

GLM is a flexible extension of ordinary linear regression that allows the response variable to follow different distributions. By combining R with Stan, GLMs can be extended into a Bayesian framework.

- **Linear model**: used for continuous outcomes with normal distribution.
- **Logit-Binomial model**: used for binary or proportion outcomes.
- **Poisson regression model**: used for count data such as event frequencies.

These analyses can also be performed using maximum likelihood estimation approaches (**link**).