# Case Study: Life Expectancy Data — Core Tasks

This handout guides you through the case study tasks. It contains prompts and hints only. Use your R script to complete the tasks.

## 0) Setup

* Ensure your working directory is set to this script's folder.
* Data file expected at: ./data/Life Expectancy Data.csv
* Read the CSV into an object named data. (Hint: use check.names = FALSE to keep original column names)
* Optional: load ggplot2 if you wish to use it.

## 1) First look at the data

* View the first few and last few rows (head(), tail()).
* Take a random sample of 10 rows (hint: sample.int()).
* Check how many rows and columns (dim(), nrow(), ncol()).
* Inspect data types and summaries (str(), summary()).
* Check for duplicate Country-Year combinations.

## 2) Data hygiene & missingness

* Count NAs per column (colSums(is.na(...))).
* Compute percentage of missing values per column (colMeans(is.na(...))\*100).
* Calculate overall missingness (mean(is.na(...))).
* Optional: Impute missing GDP values with country mean (hint: ave(), ifelse()).

## 3) Distributions & basic EDA

* Plot histograms for Life expectancy and BMI (Base R hist() or ggplot2 geom\_histogram()).
* Boxplot: Life expectancy by Status (boxplot() or ggplot2 geom\_boxplot()).
* Group summaries: mean and sd of Life expectancy by Status; median GDP; mean Schooling (tapply()).

## 4) Relationships & correlations

* Create a numeric-only data frame and compute correlation matrix (cor()).
* Compute correlations between Life expectancy and selected variables (GDP, Schooling, BMI, HIV.AIDS, Adult Mortality).
* Scatterplots: GDP vs Life expectancy, add trend line (Base R loess/lm or ggplot2 geom\_smooth()).
* Produce pairwise scatterplot matrix for selected variables (pairs()).

## 5) Programming practice

* Write a function that returns mean Life expectancy for a given country.
* Loop: for each year, print the top 3 countries by Life expectancy.
* Conditionals: create a new column lifeexp\_flag (Low <50, High >80, Typical otherwise).

## 6) Reporting-ready outputs

* Use aggregate() to produce summary tables by Status: mean Life expectancy, Schooling, GDP.
* Optional: Save a GDP vs Life expectancy plot to a file (hint: png(), plot(), lines(), dev.off()).

## 7) Reflection prompts

* Which indicators appear most strongly associated with Life Expectancy?
* Do patterns differ between Developed and Developing countries?
* How does handling missing data (e.g. imputation) affect results?
* What are the limitations of simple correlations for causal interpretation?
* If you had more time, what variables/transformations would you explore next?