# DATA WRANGLING PROJECT

Names: Chantal Gisler & Ramin Häcki

Project Title: Alcohol Consumption and Life Expectancy

━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━

## PROJECT GOAL

Objective: The goal of this project is to create a cleaned and joined dataset combining alcohol consumption patterns, smoking prevalence, and life expectancy across countries. The alcohol data comes from WHO Global Health Observatory, including recorded per capita consumption by beverage type (from 1960), unrecorded per capita consumption (from 2000, three-year averages), and drinkers-only per capita consumption (from 2000, three-year averages). Together these allow construction of total alcohol consumption, cultural drinking composition, and estimated share of drinkers. The tobacco data comes from the World Bank (2000–present), providing smoking prevalence estimates as % of adults. Life expectancy data comes from the World Bank (1960–present). These sources are messy (mixed time resolutions, missing values, metadata rows, and regional aggregates) and will be harmonized into a single reproducible dataset suitable for analysis of health and demographic relationships.

━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━

## DATA SOURCES

Primary:  
  
Source: WHO Global Health Observatory – Alcohol (https://www.who.int/data/gho/data/themes/global-information-system-on-alcohol-and-health)  
  
Description: Annual per-country data on recorded per capita consumption by beverage type (beer, wine, spirits, other) from 1960 onwards. Also includes unrecorded consumption (three-year averages from 2000) and drinkers-only consumption (three-year averages from 2000).

* Source: WHO Global Health Observatory – Alcohol (https://www.who.int/data/gho/data/themes/global-information-system-on-alcohol-and-health)
* Description: Annual per-country data on alcohol consumption per capita (beer, wine, spirits, total liters).

Secondary:  
  
Source: World Bank Life Expectancy (https://data.worldbank.org/indicator/SP.DYN.LE00.IN)  
  
Description: Country-level life expectancy at birth, annual data from 1960 onwards, including both country-specific values and regional aggregates.  
  
Additional Data:  
  
Source: World Bank Tobacco Use – Smoking Prevalence (https://data.worldbank.org/indicator/SH.PRV.SMOK)  
  
Description: Prevalence of current tobacco use (% of adults ages 15+), annual data available from 2000 onwards. Used as a control variable in combination with alcohol data when analyzing life expectancy.

* Source: World Bank Life Expectancy (https://data.worldbank.org/indicator/SP.DYN.LE00.IN)
* Description: Country-level life expectancy at birth, reported annually in years, with global coverage from the 1960s onward. The dataset includes both country-specific values and regional aggregates, with many missing values for conflict zones and small states.

━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━

## QUICK EDA

(this can also be done in separate notebook). The idea of this section is to get a high level impression of your data. E.g.:

* Shape: ~200 countries x 60 years x 5 variables (alcohol types + life expectancy)
* Missing Data: Alcohol series missing in ~15-20% of years; life expectancy missing in conflict or small island states.
* Summary Stats: Alcohol consumption range 0–15 liters per capita per year; life expectancy 40–85 years.
* Notable Patterns: Preliminary exploration suggests higher income countries show both higher alcohol intake and higher life expectancy.

━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━

## PROGRAMMING LANGUAGE

• Primary: Python

• Key Tools: Pandas, NumPy, Matplotlib, OpenPyXL

━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━

## EXPERTISE DEMONSTRATION (LE1-LE6)

How are you demonstrating expertise in:

* LE1 (Acquisition): Import WHO CSV alcohol consumption data and World Bank CSV life expectancy data.
* LE2 (Cleaning): Clean inconsistent headers, strip metadata rows, standardize country codes, impute missing alcohol years.
* LE3 (Transformation): Reshape life expectancy wide-to-long, aggregate alcohol types (beer, wine, spirits) into total liters, normalize per capita.
* LE4 (Join): Join datasets on ISO country codes and year, exclude regional aggregates.
* LE5 (Data Pipelines): Pipeline for repeatable ETL: load → clean → transform → join → output dataset.
* LE6 (Reproducibility): Reproducible Git repo, requirements.txt, documented environment, clear README.

## Analytical Timeframes

- \*\*1960–1999:\*\* Analysis can combine recorded alcohol consumption by beverage type and life expectancy. Unrecorded and drinkers-only alcohol data, as well as smoking prevalence, are not available for this period.

- \*\*2000–2020:\*\* Full analysis is possible, combining recorded consumption, unrecorded consumption, drinkers-only consumption (with derived percentage of drinkers), smoking prevalence, and life expectancy. This period allows the most complete insights.