

Life Expectancy by Region

1) Data preparation (You, 10 pts)

- How: Run `make_regions.py` to produce:
 - `data/life_expectancy_rows_full.csv` — row-level, cleaned Country×Year records (13,854 rows).
 - `data/life_expectancy_rows_sample.csv` — ~20% stratified sample by Region×Year, reproducible (rate=0.2, seed=42) (2,624 rows).
 - `life_expectancy_regions.csv` — mean life expectancy per Region×Year from the full rows.
 - `life_expectancy_regions_sample.csv` — mean life expectancy per Region×Year from the sampled rows.
- Cleaning steps:
 - Keep indicator SP.DYN.LE00.IN, join Region, drop aggregates (no Region), melt years, coerce numerics, remove NaN/±∞.
- Source: World Bank WDI (WDICSV.csv, WDIcountry.csv).

2) Apply core vis principles: choose one task (You, 15 pts)

- Chosen question: Is there a correlation between life expectancy and region over the years?
→ This is a time question, so we use a line chart with Region as series.

3) Generate a chart from the small data (AI, 5 pts)

- "Sample" button: load `life_expectancy_regions_sample.csv`.

4) Generate a chart from the large data (AI, 5 pts)

- "Full" button: load `life_expectancy_regions.csv`.

5) Check awareness of different data sizes (You, 10 pts)

- Observation: AI keeps the same code and visual style, only swapping the data source.

6) Critique the AI output with guidelines (You, 15 pts)

- Check here: <https://chatgpt.com/share/68e5bba2-ad98-8013-a0dc-11fca4477398> or `gpt_chat_history.webarchive` if the linke is deprecated.

7) Ask AI to fix items one-by-one (AI, 5 pts)

- Check here: <https://chatgpt.com/share/68e5bba2-ad98-8013-a0dc-11fca4477398> or `gpt_chat_history.webarchive` if the linke is deprecated.

8) When AI cannot fix, you fix one item (AI + You, 20 pts)

- Clear the Filter table in "Full" and "Sample".
- Change Difference to absolute value.
- Also clean up some comments and labels that may be misleading.

9) Ask AI to justify design choices (AI, 5 pts)

- Views built for the question: single (Full/Sample), Overlay (solid vs dashed, same axes), and Difference (Sample – Full) centered at 0 to expose sampling effects.
- Fair comparison: fixed x/y domains across modes so lines don't shift when toggling.
- Defensible axes: linear scales; labeled Year and Life expectancy (years); baseline at 0 (and 0-baseline for differences in years).
- Color & legend: colorblind-safe (Tableau10), color encodes only region, stable legend order/colors across all views.

- Interaction where it helps: region filters only in Overlay and Difference; disabled for Full/Sample to keep them authoritative.
- Precision on hover: tooltips show Region, Year, and exact value (or difference) in years.
- Accessibility/usability: clear button labels with aria-pressed; uncluttered layout and consistent styling.

10) What did you learn through the process? Critically evaluate (You, 10 pts)

- I learned that small design choices strongly affect inference. Keeping shared axes across Full/Sample was essential—without it, apparent gaps between regions were partly scale artifacts. Making Difference an absolute value clarified the magnitude of sampling error; however, it also hides direction, so I exposed the signed values in data/CSV to keep the analysis honest. Moving filters to Overlay/Diff only reduced cognitive load when reading baselines.