An R Markdown document converted from “Week5_Exercise2_HDI.ipynb”

Exercise 2: Human Development Index Analysis

Author: Amogh Guthur

This notebook reads HDI Table 1 data from GitHub, cleans the dataset to keep only countries without missing values, computes summary statistics, and creates visualizations.

Setup: Install and Load Packages

```
# Install required packages for data manipulation and reading Excel files
install.packages(c("tidyverse", "readxl", "httr"), quiet = TRUE, repos = "http://cran.us.r-project.org")
```

```
# Load tidyverse for data manipulation and visualization
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.6
## ✔ forcats   1.0.1     ✔ stringr   1.6.0
## ✔ ggplot2   4.0.1     ✔ tibble    3.3.0
## ✔ lubridate 1.9.4     ✔ tidyr     1.3.2
## ✔ purrr     1.2.0
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
# Load readxl for reading Excel files
library(readxl)
```

```
# Load httr for downloading files from URLs
library(httr)
```

Step 1: Download HDI Data from GitHub

```
# Define GitHub URL for HDI Excel file
url <- "https://github.com/FundamentalsAmogh/week5_hw2/raw/main/HDR25_Statistical_Annex_HDI_Table.xlsx"

# Download file to temporary location
temp <- tempfile(fileext = ".xlsx")
GET(url, write_disk(temp, overwrite = TRUE))
```

```
## Response [https://raw.githubusercontent.com/FundamentalsAmogh/week5_hw2/main/HDR25_Statistical_Annex_HDI_Tabl
e.xlsx]
##   Date: 2026-01-20 09:09
##   Status: 200
##   Content-Type: application/octet-stream
##   Size: 44.1 kB
## <ON DISK>  /var/folders/7g/hzr_s5fj7hq62fc5yhz1cp6h0000gn/T//RtmpUWgL5A/file893230fa7b.xlsx
```

Step 2: Read and Inspect the Excel File

```
# View sheet names in the Excel file
excel_sheets(temp)
```

```
## [1] "Table 1. HDI"
```

```
# Read the raw data to inspect structure (no skipping)
hdi_peek <- read_excel(temp, sheet = 1, n_max = 15)
```

```
## New names:
## • `` -> `...1`
## • `` -> `...3`
## • `` -> `...4`
## • `` -> `...5`
## • `` -> `...6`
## • `` -> `...7`
## • `` -> `...8`
## • `` -> `...9`
## • `` -> `...10`
## • `` -> `...11`
## • `` -> `...12`
## • `` -> `...13`
## • `` -> `...14`
## • `` -> `...15`
```

```
hdi_peek
```

```
## # A tibble: 15 × 15
##    ...1   Table 1. Human Devel…[1] ...3  ...4  ...5  ...6  ...7  ...8  ...9  ...10
##    <chr>  <chr>                    <chr> <lgl> <chr> <chr> <chr> <chr> <chr> <chr>
##  1 <NA>   <NA>                     <NA>  NA    <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
##  2 <NA>   <NA>                     <NA>  NA    SDG3  <NA>  SDG4… <NA>  SDG4… <NA>
##  3 <NA>   <NA>                     <NA>  NA    <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
##  4 <NA>   <NA>                     Huma… NA    Life… <NA>  Expe… <NA>  Mean… <NA>
##  5 HDI r… Country                  Value NA    (yea… <NA>  (yea… <NA>  (yea… <NA>
##  6 <NA>   <NA>                     2023  NA    2023  <NA>  2023  a     2023  a
##  7 <NA>   Very high human devel…   <NA>  NA    <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
##  8 1      Iceland                  0.97… NA    82.6… <NA>  18.8… c     13.9… d
##  9 2      Norway                   0.97  NA    83.3… <NA>  18.7… c     13.1… e
## 10 2      Switzerland              0.97  NA    83.9… <NA>  16.6… <NA>  13.9… e
## 11 4      Denmark                  0.96… NA    81.9… <NA>  18.7… c     13.0… e
## 12 5      Germany                  0.95… NA    81.3… <NA>  17.3… <NA>  14.2… e
## 13 5      Sweden                   0.95… NA    83.2… <NA>  18.9… c     12.7… e
## 14 7      Australia                0.95… NA    83.9… <NA>  20.6… c     12.8… <NA>
## 15 8      Hong Kong, China (SAR)   0.95… NA    85.5… g     16.8… <NA>  12.3… <NA>
## # ℹ abbreviated name: [1]`Table 1. Human Development Index and its components`
## # ℹ 5 more variables: ...11 <chr>, ...12 <chr>, ...13 <chr>, ...14 <chr>,
## #   ...15 <chr>
```

```
# Read the data skipping appropriate header rows
hdi_raw <- read_excel(temp, sheet = 1)
```

```
## New names:
## • `` -> `...1`
## • `` -> `...3`
## • `` -> `...4`
## • `` -> `...5`
## • `` -> `...6`
## • `` -> `...7`
## • `` -> `...8`
## • `` -> `...9`
## • `` -> `...10`
## • `` -> `...11`
## • `` -> `...12`
## • `` -> `...13`
## • `` -> `...14`
## • `` -> `...15`
```

```
# View column names
names(hdi_raw)
```

```
##  [1] "...1"
##  [2] "Table 1. Human Development Index and its components"
##  [3] "...3"
##  [4] "...4"
##  [5] "...5"
##  [6] "...6"
##  [7] "...7"
##  [8] "...8"
##  [9] "...9"
## [10] "...10"
## [11] "...11"
## [12] "...12"
## [13] "...13"
## [14] "...14"
## [15] "...15"
```

```
# View first 20 rows
head(hdi_raw, 20)
```

```
## # A tibble: 20 × 15
##    ...1   Table 1. Human Devel…¹ ...3  ...4  ...5  ...6  ...7  ...8  ...9  ...10
##    <chr>  <chr>                  <chr> <lgl> <chr> <chr> <chr> <chr> <chr> <chr>
##  1 <NA>   <NA>                   <NA>  NA    <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
##  2 <NA>   <NA>                   <NA>  NA    SDG3  <NA>  SDG4… <NA>  SDG4… <NA>
##  3 <NA>   <NA>                   <NA>  NA    <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
##  4 <NA>   <NA>                   Huma… NA    Life… <NA>  Expe… <NA>  Mean… <NA>
##  5 HDI r… Country                Value NA    (yea… <NA>  (yea… <NA>  (yea… <NA>
##  6 <NA>   <NA>                   2023  NA    2023  <NA>  2023  a     2023  a
##  7 <NA>   Very high human devel… <NA>  NA    <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
##  8 1      Iceland                0.97… NA    82.6… <NA>  18.8… c     13.9… d
##  9 2      Norway                 0.97  NA    83.3… <NA>  18.7… c     13.1… e
## 10 2      Switzerland            0.97  NA    83.9… <NA>  16.6… <NA>  13.9… e
## 11 4      Denmark                0.96… NA    81.9… <NA>  18.7… c     13.0… e
## 12 5      Germany                0.95… NA    81.3… <NA>  17.3… <NA>  14.2… e
## 13 5      Sweden                 0.95… NA    83.2… <NA>  18.9… c     12.7… e
## 14 7      Australia              0.95… NA    83.9… <NA>  20.6… c     12.8… <NA>
## 15 8      Hong Kong, China (SAR) 0.95… NA    85.5… g     16.8… <NA>  12.3… <NA>
## 16 8      Netherlands            0.95… NA    82.1… <NA>  18.5… c     12.6… e
## 17 10     Belgium                0.95… NA    82.1… <NA>  18.9… c     12.6… e
## 18 11     Ireland                0.94… NA    82.4… <NA>  19.1… c     11.7… e
## 19 12     Finland                0.94… NA    81.91 <NA>  19.4… c     12.9… e
## 20 13     Singapore              0.94… NA    83.7… <NA>  16.7… <NA>  11.9… <NA>
## # ℹ abbreviated name: ¹`Table 1. Human Development Index and its components`
## # ℹ 5 more variables: ...11 <chr>, ...12 <chr>, ...13 <chr>, ...14 <chr>,
## #   ...15 <chr>
```

Step 3: Select and Rename Columns

IMPORTANT FIX: The Excel file has alternating data columns and empty columns. Looking at the structure: - Column 1: HDI Rank - Column 2: Country
- Column 3: HDI Value - Column 4: **Empty (NA)** - Column 5: Life Expectancy - Column 6: **Empty (NA)** - Column 7: Expected Years of Schooling - Column 8: **Empty (NA)** - Column 9: Mean Years of Schooling - Column 10: **Empty (NA)** - Column 11: GNI per Capita

We need to select columns **1, 2, 3, 5, 7, 9, 11** (skipping the empty ones).

```
# Select the CORRECT columns (skipping the empty NA columns)
# Columns: 1=Rank, 2=Country, 3=HDI, 5=Life Exp, 7=Expected School, 9=Mean School, 11=GNI
hdi_select <- hdi_raw[, c(1, 2, 3, 5, 7, 9, 11)]
```

```
# Rename columns to clean names
names(hdi_select) <- c("HDI_Rank", "Country", "HDI_Value", "Life_Expectancy",
                       "Expected_Years_Schooling", "Mean_Years_Schooling", "GNI_Per_Capita")
```

```
# View renamed data
head(hdi_select, 20)
```

```
## # A tibble: 20 × 7
##    HDI_Rank Country             HDI_Value Life_Expectancy Expected_Years_Schoo…¹
##    <chr>    <chr>               <chr>     <chr>           <chr>
##  1 <NA>     <NA>                <NA>      <NA>            <NA>
##  2 <NA>     <NA>                <NA>      SDG3            SDG4.3
##  3 <NA>     <NA>                <NA>      <NA>            <NA>
##  4 <NA>     <NA>                Human De… Life expectanc… Expected years of sch…
##  5 HDI rank Country             Value     (years)         (years)
##  6 <NA>     <NA>                2023      2023            2023
##  7 <NA>     Very high human de… <NA>      <NA>            <NA>
##  8 1        Iceland             0.971999… 82.69100000000… 18.850589750000001
##  9 2        Norway              0.97      83.30800000000… 18.792850489999999
## 10 2        Switzerland         0.97      83.95399999999… 16.667530060000001
## 11 4        Denmark             0.961999… 81.93300000000… 18.704010010000001
## 12 5        Germany             0.958999… 81.378          17.30921936
## 13 5        Sweden              0.958999… 83.262          18.991470339999999
## 14 7        Australia           0.957999… 83.92300000000… 20.654779430000001
## 15 8        Hong Kong, China (… 0.954999… 85.51099999999… 16.895860670000001
## 16 8        Netherlands         0.954999… 82.15800000000… 18.58485031
## 17 10       Belgium             0.950999… 82.11499999999… 18.996030810000001
## 18 11       Ireland             0.948999… 82.41200000000… 19.184879299999999
## 19 12       Finland             0.947999… 81.91           19.494089129999999
## 20 13       Singapore           0.945999… 83.73600000000… 16.74227905
## # ℹ abbreviated name: ¹Expected_Years_Schooling
## # ℹ 2 more variables: Mean_Years_Schooling <chr>, GNI_Per_Capita <chr>
```

Step 4: Clean the Data - Keep Only Countries

```
# Remove rows where HDI_Rank is not a valid number (these are headers/notes)
hdi_clean <- hdi_select %>%
  filter(!is.na(suppressWarnings(as.numeric(HDI_Rank))))
```

```
# Convert columns to proper numeric types
hdi_clean <- hdi_clean %>% mutate(
  HDI_Rank = as.numeric(HDI_Rank),
  HDI_Value = as.numeric(HDI_Value),
  Life_Expectancy = as.numeric(Life_Expectancy),
  Expected_Years_Schooling = as.numeric(Expected_Years_Schooling),
  Mean_Years_Schooling = as.numeric(Mean_Years_Schooling),
  GNI_Per_Capita = as.numeric(GNI_Per_Capita)
)
```

```
# Check how many rows we have now
cat("Rows after filtering by valid HDI_Rank:", nrow(hdi_clean))
```

```
## Rows after filtering by valid HDI_Rank: 193
```

Step 5: Remove Countries with Missing Values

```
# Count missing values in each column
colSums(is.na(hdi_clean))
```

```
##                 HDI_Rank                  Country                HDI_Value
##                        0                        0                        0
##          Life_Expectancy Expected_Years_Schooling     Mean_Years_Schooling
##                        0                        0                        0
##           GNI_Per_Capita
##                        0
```

```
# Remove rows with any missing values
hdi_final <- na.omit(hdi_clean)
```

```
# Check final row count
cat("Final number of countries:", nrow(hdi_final))
```

```
## Final number of countries: 193
```

Step 6: Verify Cleaned Data

```
# Check structure of final data
str(hdi_final)
```

```
## tibble [193 × 7] (S3: tbl_df/tbl/data.frame)
##  $ HDI_Rank                : num [1:193] 1 2 2 4 5 5 7 8 8 10 ...
##  $ Country                 : chr [1:193] "Iceland" "Norway" "Switzerland" "Denmark" ...
##  $ HDI_Value               : num [1:193] 0.972 0.97 0.97 0.962 0.959 0.959 0.958 0.955 0.955 0.951 ...
##  $ Life_Expectancy         : num [1:193] 82.7 83.3 84 81.9 81.4 ...
##  $ Expected_Years_Schooling: num [1:193] 18.9 18.8 16.7 18.7 17.3 ...
##  $ Mean_Years_Schooling    : num [1:193] 13.9 13.1 13.9 13 14.3 ...
##  $ GNI_Per_Capita          : num [1:193] 69117 112710 81949 76008 64053 ...
```

```
# View first 15 rows
head(hdi_final, 15)
```

```
## # A tibble: 15 × 7
##    HDI_Rank Country             HDI_Value Life_Expectancy Expected_Years_Schoo…¹
##       <dbl> <chr>                   <dbl>           <dbl>                  <dbl>
##  1        1 Iceland                 0.972            82.7                   18.9
##  2        2 Norway                  0.97             83.3                   18.8
##  3        2 Switzerland             0.97             84.0                   16.7
##  4        4 Denmark                 0.962            81.9                   18.7
##  5        5 Germany                 0.959            81.4                   17.3
##  6        5 Sweden                  0.959            83.3                   19.0
##  7        7 Australia               0.958            83.9                   20.7
##  8        8 Hong Kong, China (…     0.955            85.5                   16.9
##  9        8 Netherlands             0.955            82.2                   18.6
## 10       10 Belgium                 0.951            82.1                   19.0
## 11       11 Ireland                 0.949            82.4                   19.2
## 12       12 Finland                 0.948            81.9                   19.5
## 13       13 Singapore               0.946            83.7                   16.7
## 14       13 United Kingdom          0.946            81.3                   17.8
## 15       15 United Arab Emirat…     0.94             82.9                   15.6
## # ℹ abbreviated name: ¹Expected_Years_Schooling
## # ℹ 2 more variables: Mean_Years_Schooling <dbl>, GNI_Per_Capita <dbl>
```

```
# View last 10 rows
tail(hdi_final, 10)
```

```
## # A tibble: 10 × 7
##    HDI_Rank Country             HDI_Value Life_Expectancy Expected_Years_Schoo…¹
##       <dbl> <chr>                   <dbl>           <dbl>                  <dbl>
##  1      184 Yemen                   0.47             69.3                   7.49
##  2      185 Sierra Leone            0.467            61.8                   9.06
##  3      186 Burkina Faso            0.459            61.1                   8.73
##  4      187 Burundi                 0.439            63.7                   9.83
##  5      188 Mali                    0.419            60.4                   7.01
##  6      188 Niger                   0.419            61.2                   8.31
##  7      190 Chad                    0.416            55.1                   8.35
##  8      191 Central African Re…     0.414            57.4                   7.44
##  9      192 Somalia                 0.404            58.8                   7.49
## 10      193 South Sudan             0.388            57.6                   5.63
## # ℹ abbreviated name: ¹Expected_Years_Schooling
## # ℹ 2 more variables: Mean_Years_Schooling <dbl>, GNI_Per_Capita <dbl>
```

```
# Summary statistics
summary(hdi_final)
```

```
##     HDI_Rank       Country           HDI_Value      Life_Expectancy
##  Min.   :  1.0   Length:193         Min.   :0.3880   Min.   :54.46
##  1st Qu.: 48.0   Class :character   1st Qu.:0.6220   1st Qu.:67.39
##  Median : 97.0   Mode  :character   Median :0.7620   Median :73.49
##  Mean   : 96.8                      Mean   :0.7408   Mean   :73.11
##  3rd Qu.:145.0                      3rd Qu.:0.8620   3rd Qu.:78.34
##  Max.   :193.0                      Max.   :0.9720   Max.   :85.71
##  Expected_Years_Schooling Mean_Years_Schooling GNI_Per_Capita
##  Min.   : 5.635           Min.   : 1.412       Min.   :   688.3
##  1st Qu.:11.505           1st Qu.: 6.780       1st Qu.:  5746.6
##  Median :13.336           Median : 9.933       Median : 15866.5
##  Mean   :13.585           Mean   : 9.173       Mean   : 24620.7
##  3rd Qu.:15.888           3rd Qu.:11.642       3rd Qu.: 36793.0
##  Max.   :20.846           Max.   :14.296       Max.   :166811.7
```

Step 7: Compute Mean of Key Variables

As required by the assignment, computing the mean of: Life expectancy at birth, Expected years of schooling, Mean years of schooling, and GNI per capita.

```
# Compute mean of Life Expectancy at birth
mean_life_exp <- mean(hdi_final$Life_Expectancy, na.rm = TRUE)
cat("Mean Life Expectancy at birth:", round(mean_life_exp, 2), "years")
```

```
## Mean Life Expectancy at birth: 73.11 years
```

```
# Compute mean of Expected Years of Schooling
mean_expected_school <- mean(hdi_final$Expected_Years_Schooling, na.rm = TRUE)
cat("Mean Expected Years of Schooling:", round(mean_expected_school, 2), "years")
```

```
## Mean Expected Years of Schooling: 13.58 years
```

```
# Compute mean of Mean Years of Schooling
mean_years_school <- mean(hdi_final$Mean_Years_Schooling, na.rm = TRUE)
cat("Mean Years of Schooling:", round(mean_years_school, 2), "years")
```

```
## Mean Years of Schooling: 9.17 years
```

```
# Compute mean of GNI per Capita
mean_gni <- mean(hdi_final$GNI_Per_Capita, na.rm = TRUE)
cat("Mean GNI per Capita:", round(mean_gni, 2), "(2021 PPP $)")
```

```
## Mean GNI per Capita: 24620.68 (2021 PPP $)
```

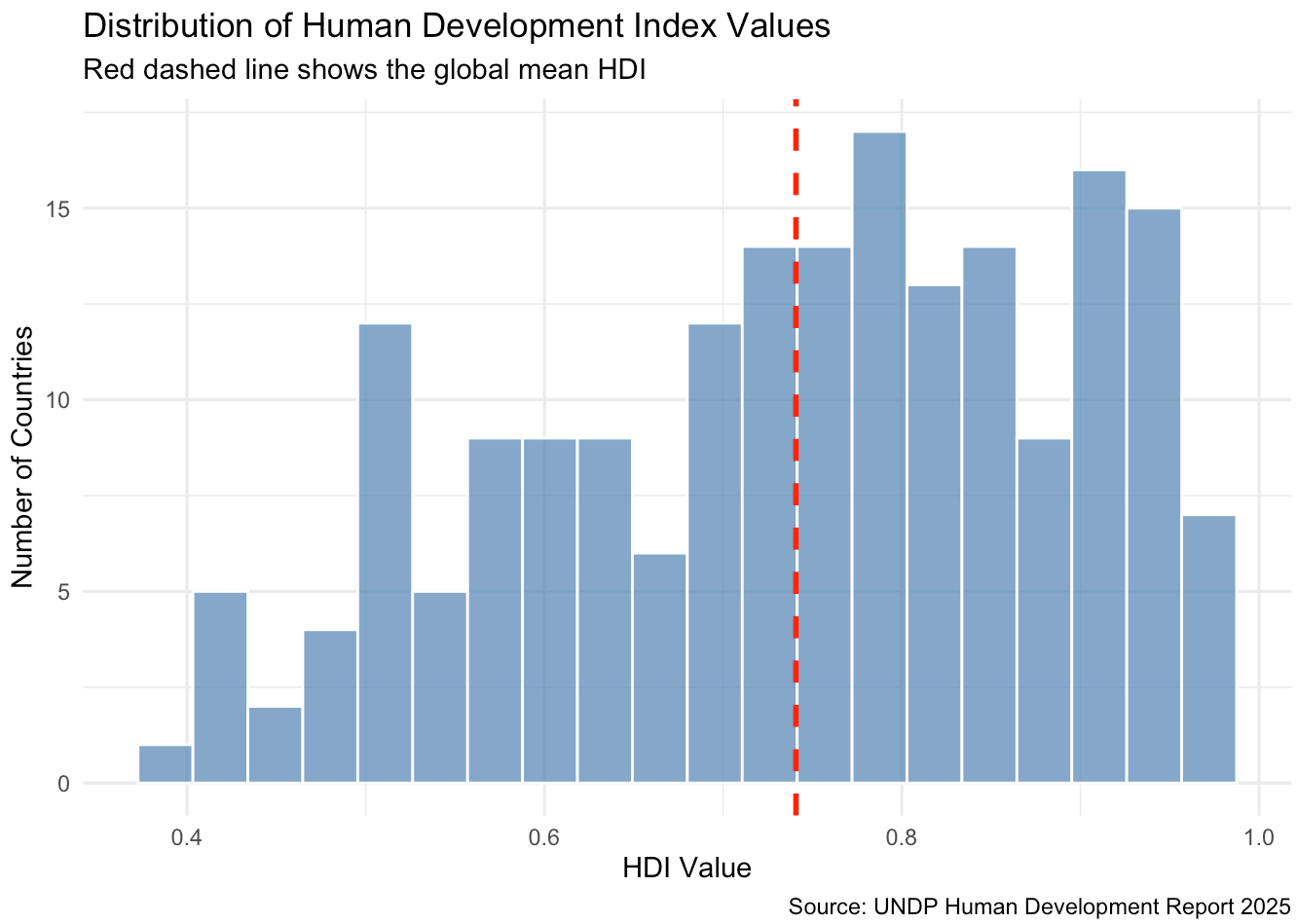
```
# Display all means in a summary table
means_summary <- data.frame(
  Variable = c("Life Expectancy at birth", "Expected Years of Schooling",
               "Mean Years of Schooling", "GNI per Capita"),
  Mean = c(round(mean_life_exp, 2), round(mean_expected_school, 2),
           round(mean_years_school, 2), round(mean_gni, 2)),
  Unit = c("years", "years", "years", "2021 PPP $")
)
means_summary
```

```
##                      Variable     Mean       Unit
## 1    Life Expectancy at birth    73.11      years
## 2 Expected Years of Schooling    13.58      years
## 3     Mean Years of Schooling     9.17      years
## 4              GNI per Capita 24620.68 2021 PPP $
```

Step 8: Visualization 1 - HDI Value Distribution

This histogram shows the distribution of Human Development Index values across all countries. The HDI ranges from 0 to 1, with higher values indicating better human development. The red dashed line shows the global mean HDI.

```
# Create histogram of HDI values showing distribution across countries
ggplot(hdi_final, aes(x = HDI_Value)) +
  geom_histogram(bins = 20, fill = "steelblue", color = "white", alpha = 0.7) +
  geom_vline(xintercept = mean(hdi_final$HDI_Value, na.rm = TRUE),
             color = "red", linetype = "dashed", linewidth = 1) +
  labs(title = "Distribution of Human Development Index Values",
       subtitle = "Red dashed line shows the global mean HDI",
       x = "HDI Value",
       y = "Number of Countries",
       caption = "Source: UNDP Human Development Report 2025") +
  theme_minimal()
```

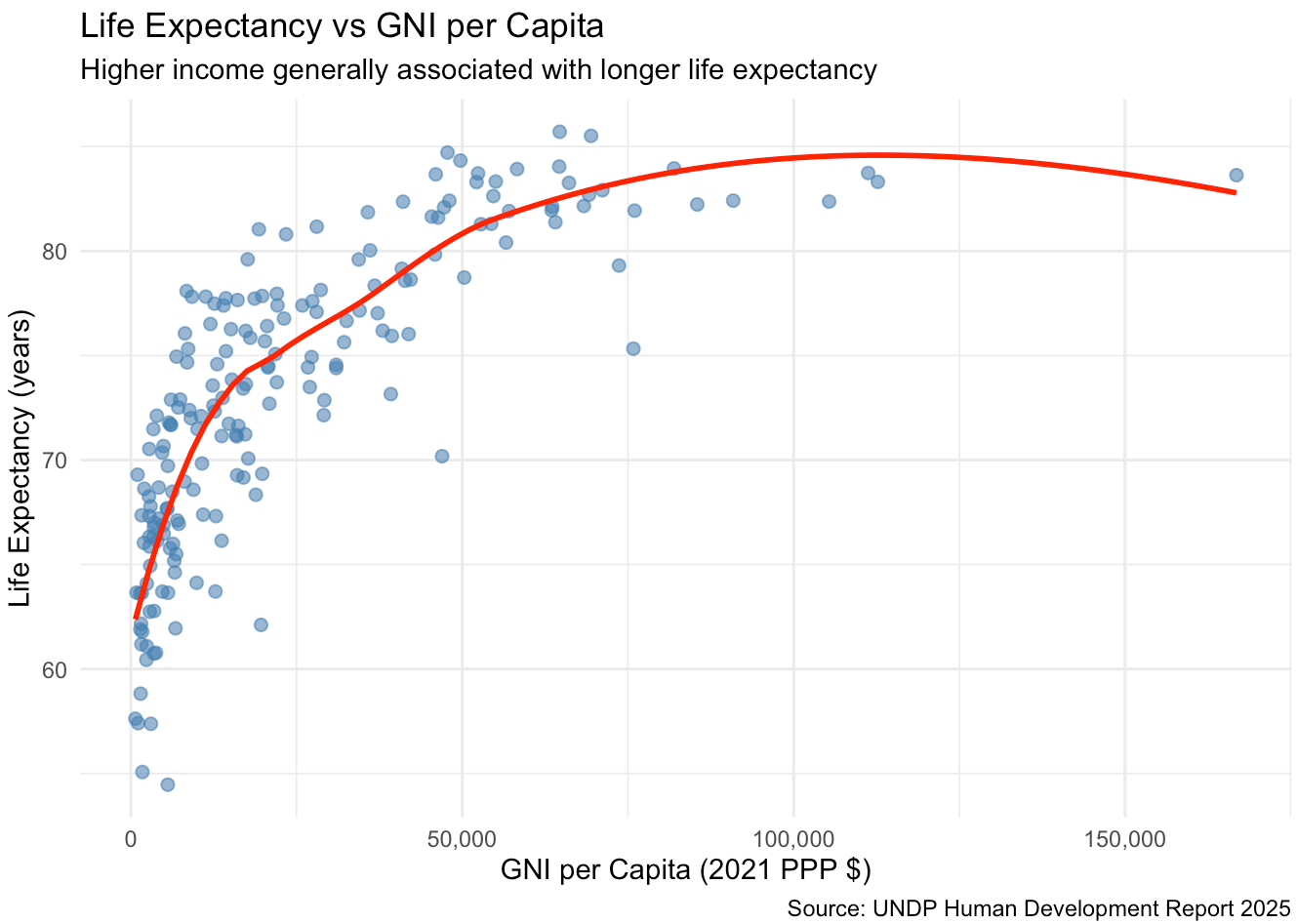


Step 9: Visualization 2 - Life Expectancy vs GNI per Capita

This scatter plot shows the relationship between Life Expectancy and GNI per Capita. These are two of the three dimensions used to calculate HDI. The plot reveals that higher income countries generally have higher life expectancy, though the relationship shows diminishing returns at higher income levels.

```
# Create scatter plot showing relationship between life expectancy and income
ggplot(hdi_final, aes(x = GNI_Per_Capita, y = Life_Expectancy)) +
  geom_point(alpha = 0.6, color = "steelblue", size = 2) +
  geom_smooth(method = "loess", color = "red", se = FALSE) +
  scale_x_continuous(labels = scales::comma) +
  labs(title = "Life Expectancy vs GNI per Capita",
       subtitle = "Higher income generally associated with longer life expectancy",
       x = "GNI per Capita (2021 PPP $)",
       y = "Life Expectancy (years)",
       caption = "Source: UNDP Human Development Report 2025") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Step 10: Visualization 3 - Education Indicators Comparison

This scatter plot compares Expected Years of Schooling with Mean Years of Schooling. Expected years represents future educational attainment for children entering school, while mean years shows current adult education levels. Points above the diagonal line indicate countries where future generations are expected to receive more education than current adults.

```
# Create scatter plot comparing education indicators
ggplot(hdi_final, aes(x = Mean_Years_Schooling, y = Expected_Years_Schooling)) +
  geom_point(alpha = 0.6, color = "steelblue", size = 2) +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "gray50") +
  labs(title = "Expected vs Mean Years of Schooling",
       subtitle = "Points above diagonal indicate improving educational attainment",
       x = "Mean Years of Schooling (adults)",
       y = "Expected Years of Schooling (children)",
       caption = "Source: UNDP Human Development Report 2025") +
  theme_minimal()
```