WHO Life Expectancy — Cleaning & EDA

Dominion Samuel

August 26,2025

Introduction

This notebook focuses on cleaning and exploring the Life Expectancy dataset. The dataset contains global health indicators over multiple years for developing and developed countries.

My aim is to prepare a clean dataset for later statistical modeling, focusing on developing countries from the year 2000 onward. I also examine variable distributions and correlations to identify potential predictors of low life expectancy.

Download and load data

```
# Define URLs and paths
raw_url <- "https://raw.githubusercontent.com/selva86/datasets/master/Life_Expectancy_Data.csv"
dest    <- "../data/Life_Expectancy_Data.csv"

# Download dataset if not already present
if (!file.exists(dest)) {
  download.file(raw_url, destfile = dest, mode = "wb")
}
# Load dataset and clean column names
raw <- read_csv(dest) |> clean_names()
```

```
## Rows: 1649 Columns: 22
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): Country, Status
## dbl (20): Year, Life expectancy, Adult Mortality, infant deaths, Alcohol, pe...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
glimpse(raw)
```

```
## Rows: 1,649
## Columns: 22
## $ country                          <chr> "Afghanistan", "Afghanistan", "Afghani~
## $ year                             <dbl> 2015, 2014, 2013, 2012, 2011, 2010, 20~
## $ status                           <chr> "Developing", "Developing", "Developin~
```

```
## $ life_expectancy                <dbl> 65.0, 59.9, 59.9, 59.5, 59.2, 58.8, 58~
## $ adult_mortality                <dbl> 263, 271, 268, 272, 275, 279, 281, 287~
## $ infant_deaths                  <dbl> 62, 64, 66, 69, 71, 74, 77, 80, 82, 84~
## $ alcohol                        <dbl> 0.01, 0.01, 0.01, 0.01, 0.01, 0.01, 0.~
## $ percentage_expenditure         <dbl> 71.279624, 73.523582, 73.219243, 78.18~
## $ hepatitis_b                    <dbl> 65, 62, 64, 67, 68, 66, 63, 64, 63, 64~
## $ measles                        <dbl> 1154, 492, 430, 2787, 3013, 1989, 2861~
## $ bmi                            <dbl> 19.1, 18.6, 18.1, 17.6, 17.2, 16.7, 16~
## $ under_five_deaths              <dbl> 83, 86, 89, 93, 97, 102, 106, 110, 113~
## $ polio                          <dbl> 6, 58, 62, 67, 68, 66, 63, 64, 63, 58,~
## $ total_expenditure              <dbl> 8.16, 8.18, 8.13, 8.52, 7.87, 9.20, 9.~
## $ diphtheria                     <dbl> 65, 62, 64, 67, 68, 66, 63, 64, 63, 58~
## $ hiv_aids                       <dbl> 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1~
## $ gdp                            <dbl> 584.25921, 612.69651, 631.74498, 669.9~
## $ population                     <dbl> 33736494, 327582, 31731688, 3696958, 2~
## $ thinness_1_19_years            <dbl> 17.2, 17.5, 17.7, 17.9, 18.2, 18.4, 18~
## $ thinness_5_9_years             <dbl> 17.3, 17.5, 17.7, 18.0, 18.2, 18.4, 18~
## $ income_composition_of_resources <dbl> 0.479, 0.476, 0.470, 0.463, 0.454, 0.4~
## $ schooling                      <dbl> 10.1, 10.0, 9.9, 9.8, 9.5, 9.2, 8.9, 8~
```

```
skim(raw)
```

Table 1: Data summary

Name	raw
Number of rows	1649
Number of columns	22
Column type frequency:	
character	2
numeric	20
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
country	0	1	4	24	0	133	0
status	0	1	9	10	0	2	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
year	0	1	2007.84	4.09	2000.00	2005.00	2008.00	2011.00	2.015000e+03	
life_expectancy	0	1	69.30	8.80	44.00	64.40	71.70	75.00	8.900000e+01	
adult_mortality	0	1	168.22	125.31	1.00	77.00	148.00	227.00	7.230000e+02	
infant_deaths	0	1	32.55	120.85	0.00	1.00	3.00	22.00	1.600000e+03	
alcohol	0	1	4.53	4.03	0.01	0.81	3.79	7.34	1.787000e+01	
percentage_expenditure	0	1	698.97	1759.23	0.00	37.44	145.10	509.39	1.896135e+04	
hepatitis_b	0	1	79.22	25.60	2.00	74.00	89.00	96.00	9.900000e+01	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
measles	0	1	2224.49	10085.80	0.00	0.00	15.00	373.00	1.314410e+05	
bmi	0	1	38.13	19.75	2.00	19.50	43.70	55.80	7.710000e+01	
under_five_deaths	0	1	44.22	162.90	0.00	1.00	4.00	29.00	2.100000e+03	
polio	0	1	83.56	22.45	3.00	81.00	93.00	97.00	9.900000e+01	
total_expenditure	0	1	5.96	2.30	0.74	4.41	5.84	7.47	1.439000e+01	
diphtheria	0	1	84.16	21.58	2.00	82.00	92.00	97.00	9.900000e+01	
hiv_aids	0	1	1.98	6.03	0.10	0.10	0.10	0.70	5.060000e+01	
gdp	0	1	5566.03	11475.90	1.68	462.15	1592.57	4718.51	1.191727e+05	
population	0	1	14653625.80	60460393.40	34.00	191897.00	1419631.00	7658972.00	1.293859e+09	
thinness_1_19_years	0	1	4.85	4.60	0.10	1.60	3.00	7.10	2.720000e+01	
thinness_5_9_years	0	1	4.91	4.65	0.10	1.70	3.20	7.10	2.820000e+01	
income_composition_of_resources	0	1	0.63	0.18	0.00	0.51	0.67	0.75	9.400000e-01	
schooling	0	1	12.12	2.80	4.20	10.30	12.30	14.00	2.070000e+01	

Data Cleaning

```
# Select only Developing Countries from this century(up to 2015)
df_clean <- raw %>%
  filter(status == 'Developing', year >=2000, year <=2015)
```

Filtering ensures the analysis focuses on relevant countries and recent years.

Select Relevant Variables

```
# Select variables to focus on
df_clean <- df_clean %>%
  select(life_expectancy, adult_mortality, infant_deaths, alcohol,hepatitis_b,
         measles,bmi,polio, total_expenditure,diphtheria, hiv_aids, gdp,
         population,thinness_1_19_years, schooling )
glimpse(df_clean)
```

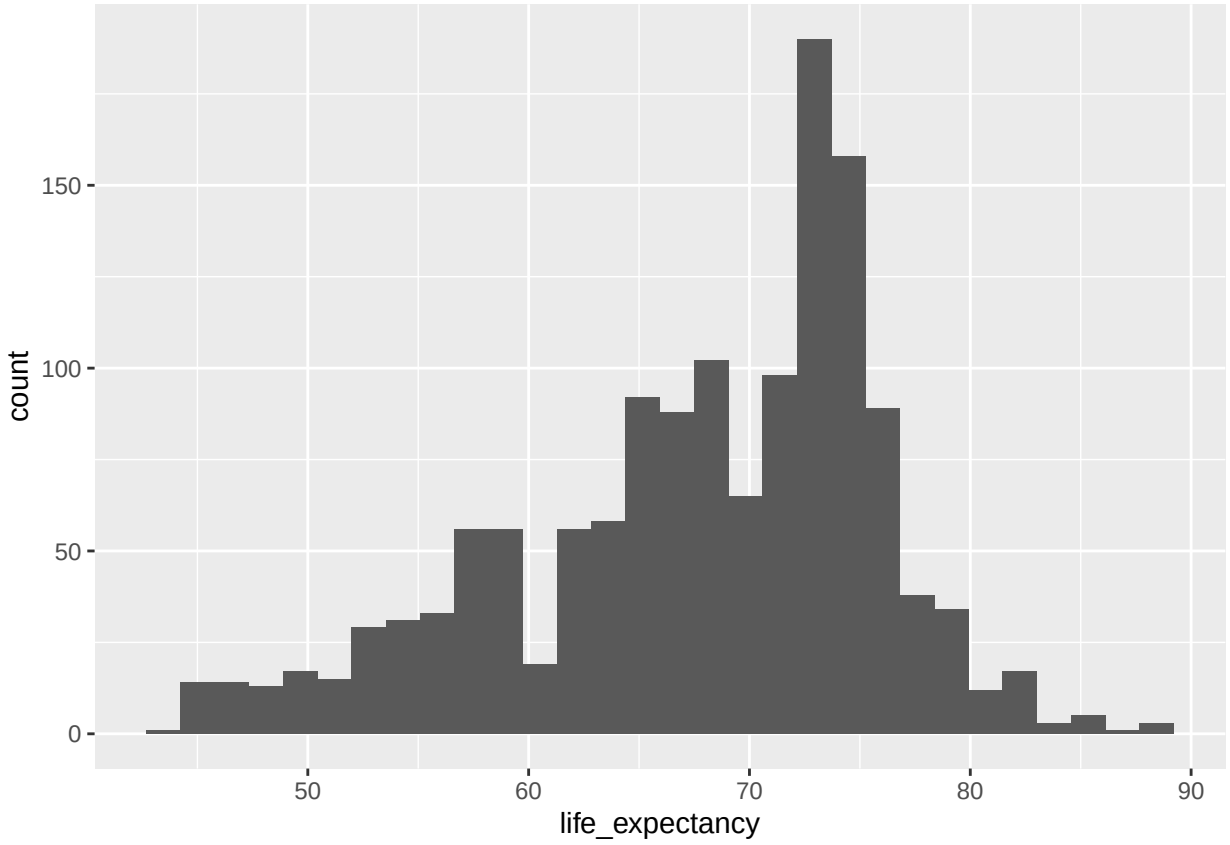
```
## Rows: 1,407
## Columns: 15
## $ life_expectancy     <dbl> 65.0, 59.9, 59.9, 59.5, 59.2, 58.8, 58.6, 58.1, 57~
## $ adult_mortality     <dbl> 263, 271, 268, 272, 275, 279, 281, 287, 295, 295, ~
## $ infant_deaths       <dbl> 62, 64, 66, 69, 71, 74, 77, 80, 82, 84, 85, 87, 87~
## $ alcohol             <dbl> 0.01, 0.01, 0.01, 0.01, 0.01, 0.01, 0.01, 0.03, 0.~
## $ hepatitis_b         <dbl> 65, 62, 64, 67, 68, 66, 63, 64, 63, 64, 66, 67, 65~
## $ measles             <dbl> 1154, 492, 430, 2787, 3013, 1989, 2861, 1599, 1141~
## $ bmi                 <dbl> 19.1, 18.6, 18.1, 17.6, 17.2, 16.7, 16.2, 15.7, 15~
## $ polio               <dbl> 6, 58, 62, 67, 68, 66, 63, 64, 63, 58, 58, 5, 41, ~
## $ total_expenditure   <dbl> 8.16, 8.18, 8.13, 8.52, 7.87, 9.20, 9.42, 8.33, 6.~
## $ diphtheria          <dbl> 65, 62, 64, 67, 68, 66, 63, 64, 63, 58, 58, 5, 41,~
## $ hiv_aids            <dbl> 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, ~
## $ gdp                 <dbl> 584.25921, 612.69651, 631.74498, 669.95900, 63.537~
## $ population          <dbl> 33736494, 327582, 31731688, 3696958, 2978599, 2883~
```

```
## $ thinness_1_19_years <dbl> 17.2, 17.5, 17.7, 17.9, 18.2, 18.4, 18.6, 18.8, 19~
## $ schooling           <dbl> 10.1, 10.0, 9.9, 9.8, 9.5, 9.2, 8.9, 8.7, 8.4, 8.1~
```

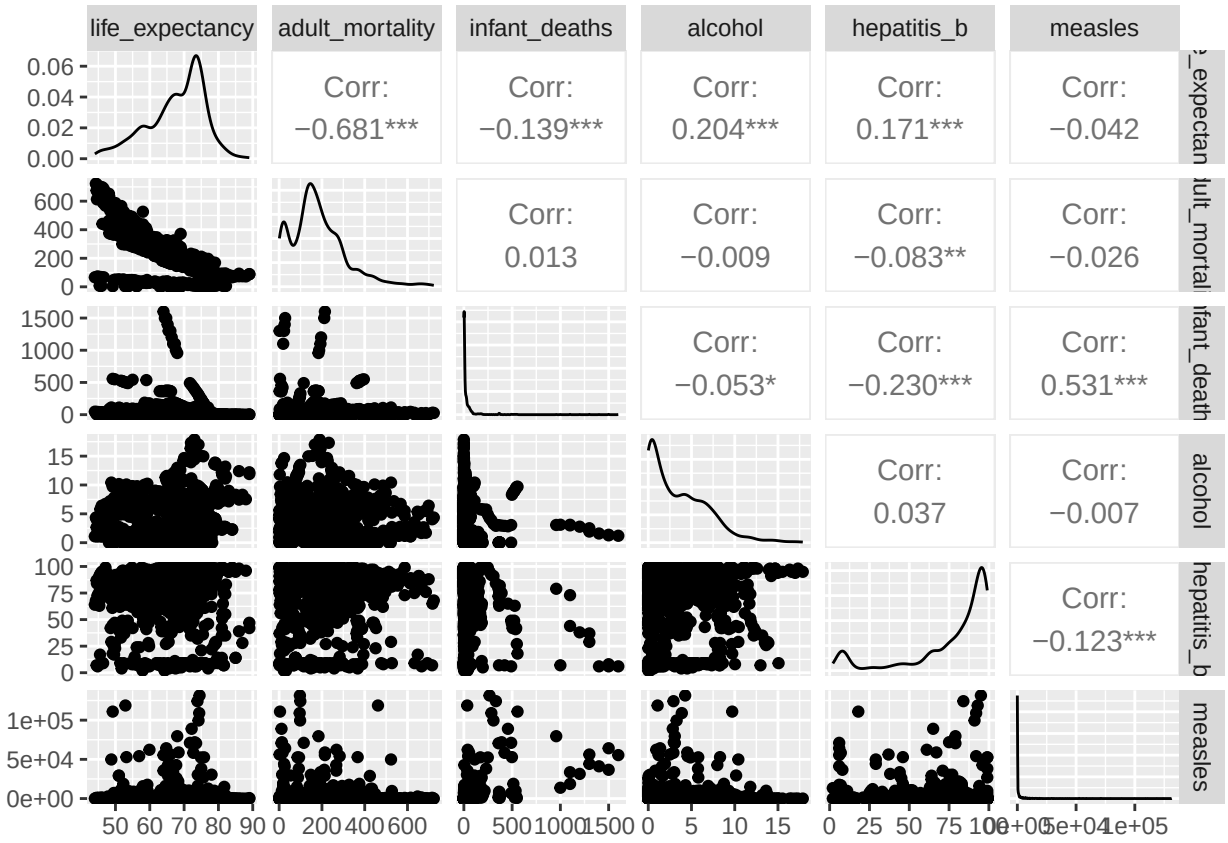
We focus on variables likely to influence life expectancy, including health indicators, disease prevalence, economic factors, and nutrition metrics.

Quick EDA

```
ggplot(df_clean, aes(life_expectancy)) + geom_histogram(bins = 30)
```




```
num_only <- df_clean |> select(where(is.numeric))
ggpairs(num_only[, 1:min(6, ncol(num_only))], progress = FALSE)
```



Save Cleaned Data

```
write_csv(df_clean, "../data/life_clean.csv")
```