STAT232_PJT

Group MESS Mehr Un Nisa Tariq Eunyoung Kwak Sahana RajeshLal
Subbareddy Bhumireddy Venkata

2024-03-15

1. Introduction

Life expectancy, reflects the average number of years a person born today is expected to live. It's a powerful indicator of a population's health, influenced by various factors like access to healthcare, nutrition, and environmental conditions. Understanding life expectancy trends across countries and over time provides valuable insights into societal progress and areas where improvements are needed.

1.1 Background of the data We found our dataset on Kaggle under the name "Life Expectancy Data". It has 2848 rows and 23 columns. It has data sourced from Global Health Observatory (GHO) data repository under World Health Organization (WHO). Our dataset has 16 years worth of data for 178 countries. For our project, we are aiming to predict life expectancy using predictors such as health-related factors, including immunization coverage for diseases like Hepatitis B, Polio, and Diphtheria, alongside mortality rates, economic indicators, and social determinants.

Loading the dataset

```
led <- read.csv("Life Expectancy Data.csv") ## Life Expectancy DataSet
countries <- read.csv("countries.csv")      ## Latitude and longitude information by country

dim(led)
```

```
## [1] 2848   23
```

```
led %>%
group_by(Region, Country, Year) %>%
  summarise(cnt = n(), .groups="drop")
```

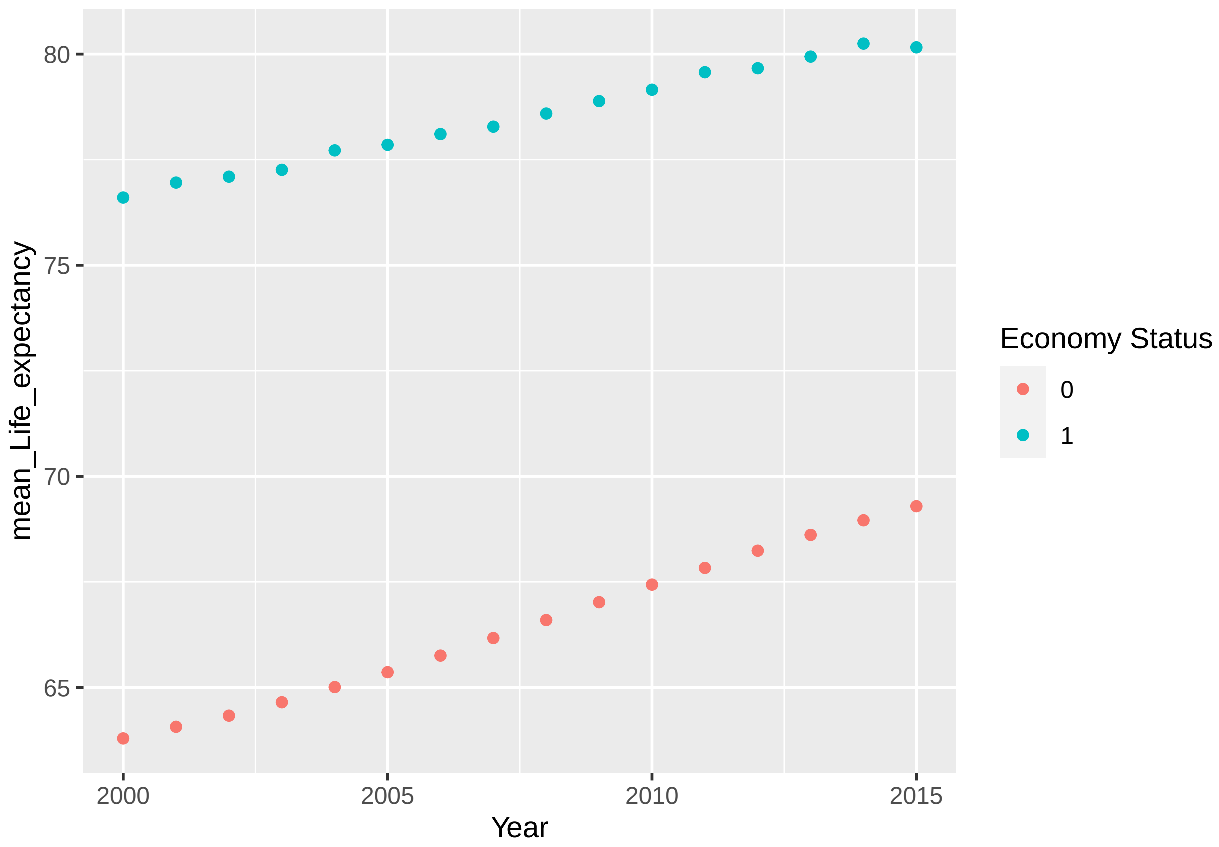
```
## # A tibble: 2,848 x 4
##    Region Country  Year   cnt
##    <chr>  <chr>   <int> <int>
##  1 Africa Algeria  2000     1
##  2 Africa Algeria  2001     1
##  3 Africa Algeria  2002     1
##  4 Africa Algeria  2003     1
##  5 Africa Algeria  2004     1
##  6 Africa Algeria  2005     1
##  7 Africa Algeria  2006     1
##  8 Africa Algeria  2007     1
##  9 Africa Algeria  2008     1
## 10 Africa Algeria  2009     1
## # i 2,838 more rows
```

1.2 variables of interest Initially, our dataset comprises of 22 predictor variables. However, our objective is to streamline this number through exploratory data analysis, correlation analysis, and the forward selection method. Our predictor names, its type and glimpse of the data is shown below.

```r
glimpse(led)
```

```
## Rows: 2,848
## Columns: 23
## $ Country                        <chr> "Afghanistan", "Afghanistan", "Afghani~
## $ Region                         <chr> "Asia", "Asia", "Asia", "Asia", "Asia"~
## $ Year                           <int> 2000, 2001, 2002, 2003, 2004, 2005, 20~
## $ Economy_status_Developed       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ Life_expectancy                <dbl> 55.8, 56.3, 56.8, 57.3, 57.8, 58.3, 58~
## $ Adult_mortality                <dbl> 310.8305, 304.8580, 298.8855, 292.0365~
## $ Infant_deaths                  <dbl> 90.5, 87.9, 85.3, 82.7, 80.0, 77.3, 74~
## $ Alcohol_consumption            <dbl> 0.020, 0.020, 0.020, 0.020, 0.020, 0.0~
## $ percentage.expenditure         <dbl> 10.424960, 10.574728, 16.887351, 11.08~
## $ Hepatitis_B                    <int> 62, 63, 64, 65, 67, 66, 64, 63, 64, 63~
## $ Measles                        <int> 6532, 8762, 2486, 798, 466, 1296, 1990~
## $ BMI                            <dbl> 12.2, 12.6, 13.0, 13.4, 13.8, 14.2, 14~
## $ Under_five_deaths              <dbl> 122, 122, 122, 122, 120, 118, 116, 113~
## $ Polio                          <int> 24, 35, 36, 41, 5, 58, 58, 63, 64, 63,~
## $ Total.expenditure              <dbl> 8.20, 7.80, 7.76, 8.82, 8.79, 8.70, 7.~
## $ Income.composition.of.resources <dbl> 0.338, 0.340, 0.341, 0.373, 0.381, 0.3~
## $ Diphtheria                     <int> 24, 33, 36, 41, 50, 58, 58, 63, 64, 63~
## $ Incidents_HIV                  <dbl> 0.02, 0.02, 0.02, 0.02, 0.02, 0.02, 0.~
## $ GDP_per_capita                 <int> 148, 163, 320, 332, 323, 346, 354, 393~
## $ Population_mln                 <dbl> 20.78, 21.61, 22.60, 23.68, 24.73, 25.~
## $ Thinness_ten_nineteen_years    <dbl> 2.3, 2.1, 19.9, 19.7, 19.5, 19.3, 19.2~
## $ Thinness_five_nine_years       <dbl> 2.5, 2.4, 2.2, 19.9, 19.7, 19.5, 19.3,~
## $ Schooling                      <dbl> 2.2, 2.2, 2.3, 2.4, 2.5, 2.6, 2.7, 2.9~
```

```r
# Assuming 'led' is the name of your dataset
led_year <- led %>%
  dplyr::select(Life_expectancy, Year, Economy_status_Developed) %>%
  group_by(Year, Economy_status_Developed) %>%
  summarise(mean_Life_expectancy = mean(Life_expectancy), .groups = "drop")

# For plotting
ggplot(data = led_year) +
  geom_point(aes(x = Year, y = mean_Life_expectancy, color = as.factor(Economy_status_Developed))) +
  labs(color = "Economy Status")
```



1.3 business questions can be answered using these data Some of the business questions we can answer from this dataset are:

- Is there a relationship between the predictors (such as "Economy_status_Developed", "Adult_mortality", "Alcohol_consumption", etc.) and life expectancy?
- How strong is the relationship between life expectancy and other predictors?

2. Exploratory Data Analysis

2.1 Data cleaning (missing data and anomaly removal, summary and aggregation)

- For better results, we are only going to use one year data since it can be challenging to deal with 16 years worth of data.
- The most recent data is from 2015, but we decided to use 2014 data because some data was missing.

```
led <- filter(led, Year==2014)
head(led,5)
```

```
##               Country                        Region Year
## 1         Afghanistan                          Asia 2014
## 2             Albania                Rest of Europe 2014
## 3             Algeria                        Africa 2014
## 4              Angola                        Africa 2014
## 5 Antigua and Barbuda Central America and Caribbean 2014
##   Economy_status_Developed Life_expectancy Adult_mortality Infant_deaths
## 1                        0            63.0        231.9780          55.2
## 2                        0            77.8         76.7240           8.8
## 3                        0            75.9         97.2770          21.9
## 4                        0            58.8        246.0945          60.5
## 5                        0            76.3        131.4520           6.8
##   Alcohol_consumption percentage.expenditure Hepatitis_B Measles  BMI
## 1                0.01               73.52358          62     492 18.6
## 2                4.10              428.74907          98       0 57.2
## 3                0.54               54.23732          95       0 58.4
## 4                8.10               23.96561          64   11699 22.7
## 5                9.66             2422.99977          99       0 47.0
##   Under_five_deaths Polio Total.expenditure Income.composition.of.resources
## 1                86    58              8.18                           0.476
## 2                 1    98              5.88                           0.761
## 3                24    95              7.21                           0.741
## 4               101    68              3.31                           0.527
## 5                 0    96              5.54                           0.782
##   Diphtheria Incidents_HIV GDP_per_capita Population_mln
## 1         62          0.03            565          33.37
## 2         98          0.03           3856           2.89
## 3         95          0.04           4112          38.92
## 4         64          0.97           3207          26.94
## 5         99          0.20          13909           0.09
##   Thinness_ten_nineteen_years Thinness_five_nine_years Schooling
## 1                        17.5                     17.5       3.5
## 2                         1.2                      1.3       9.7
## 3                         6.0                      5.8       7.9
## 4                         8.5                      8.3       4.9
## 5                         3.3                      3.3       9.2
```

- Check for missing values in the data.

```r
which(is.na(led))
```

```
## integer(0)
```

- Drop unnecessary columns: drop the Year variable because it only has 2014 as a value.

```r
led <- dplyr::select(led, -Year)
head(led, 5)
```

```
##               Country                        Region Economy_status_Developed
## 1         Afghanistan                          Asia                        0
## 2             Albania                Rest of Europe                        0
## 3             Algeria                        Africa                        0
## 4              Angola                        Africa                        0
## 5 Antigua and Barbuda Central America and Caribbean                        0
##   Life_expectancy Adult_mortality Infant_deaths Alcohol_consumption
## 1            63.0        231.9780          55.2                0.01
## 2            77.8         76.7240           8.8                4.10
## 3            75.9         97.2770          21.9                0.54
## 4            58.8        246.0945          60.5                 8.10
## 5            76.3        131.4520           6.8                 9.66
##   percentage.expenditure Hepatitis_B Measles  BMI Under_five_deaths Polio
## 1               73.52358          62     492 18.6                86    58
## 2              428.74907          98       0 57.2                 1    98
## 3               54.23732          95       0 58.4                24    95
## 4               23.96561          64   11699 22.7               101    68
## 5             2422.99977          99       0 47.0                 0    96
##   Total.expenditure Income.composition.of.resources Diphtheria Incidents_HIV
## 1              8.18                           0.476         62          0.03
## 2              5.88                           0.761         98          0.03
## 3              7.21                           0.741         95          0.04
## 4              3.31                           0.527         64          0.97
## 5              5.54                           0.782         99          0.20
##   GDP_per_capita Population_mln Thinness_ten_nineteen_years
## 1            565          33.37                        17.5
## 2           3856           2.89                         1.2
## 3           4112          38.92                         6.0
## 4           3207          26.94                         8.5
## 5          13909           0.09                         3.3
##   Thinness_five_nine_years Schooling
## 1                     17.5       3.5
## 2                      1.3       9.7
## 3                      5.8       7.9
## 4                      8.3       4.9
## 5                      3.3       9.2
```

- Join the “Life Expectancy DataSet” and `countries` datasets to get the longitude and latitude of the countries.
  - Before joining, we checked that the names of the countries in the two datasets match.

```
setdiff(led$Country, countries$name)
```

```
##  [1] "Bahamas, The"                 "Brunei Darussalam"
##  [3] "Cabo Verde"                   "Congo, Dem. Rep."
##  [5] "Congo, Rep."                  "Czechia"
##  [7] "Egypt, Arab Rep."             "Eswatini"
##  [9] "Gambia, The"                  "Iran, Islamic Rep."
## [11] "Kyrgyz Republic"              "Lao PDR"
## [13] "Micronesia, Fed. Sts."        "Myanmar"
## [15] "North Macedonia"              "Russian Federation"
## [17] "Sao Tome and Principe"        "Slovak Republic"
## [19] "St. Lucia"                    "St. Vincent and the Grenadines"
## [21] "Syrian Arab Republic"         "Turkiye"
## [23] "Venezuela, RB"                "Yemen, Rep."
```

- If the names don’t match in both datasets, rename one.

```
led[which(grepl("^Bahamas", led$Country)), 1] <- "Bahamas"
led[which(grepl("^Brunei", led$Country)), 1] <- "Brunei"
countries[which(grepl("^Cape Verde", countries$name)), 4] <- "Cabo Verde"
led[which(grepl("^Congo, Dem.", led$Country)), 1] <- "Congo [DRC]"
```

```
led[which(grepl("^Congo, Rep.", led$Country)), 1] <- "Congo [Republic]"
led[which(grepl("^Czechia", led$Country)), 1] <- "Czech Republic"
led[which(grepl("^Egypt", led$Country)), 1] <- "Egypt"
countries[which(grepl("^Swaziland", countries$name)), 4] <- "Eswatini"
led[which(grepl("^Gambia", led$Country)), 1] <- "Gambia"
led[which(grepl("^Iran", led$Country)), 1] <- "Iran"
led[which(grepl("^Kyrgyz", led$Country)), 1] <- "Kyrgyzstan"
led[which(grepl("^Lao", led$Country)), 1] <- "Laos"
led[which(grepl("^Micronesia", led$Country)), 1] <- "Micronesia"
countries[which(grepl("^Myanmar", countries$name)), 4] <- "Myanmar"
countries[which(grepl("^Macedonia", countries$name)), 4] <- "North Macedonia"
led[which(grepl("^Russia", led$Country)), 1] <- "Russia"
countries[which(grepl("^São", countries$name)), 4] <- "Sao Tome and Principe"
led[which(grepl("^Slovak", led$Country)), 1] <- "Slovakia"
led[which(grepl("^St. Lucia", led$Country)), 1] <- "Saint Lucia"
led[which(grepl("^St. Vincent", led$Country)), 1] <- "Saint Vincent and the Grenadines"
led[which(grepl("^Syrian Arab Republic", led$Country)), 1] <- "Syria"
countries[which(grepl("^Turkey", countries$name)), 4] <- "Turkiye"
led[which(grepl("^Venezuela", led$Country)), 1] <- "Venezuela"
led[which(grepl("^Yemen", led$Country)), 1] <- "Yemen"
```

- After the process, we re-checked that the names of the countries in the two datasets match.

```
setdiff(led$Country, countries$name)
```

```
## character(0)
```

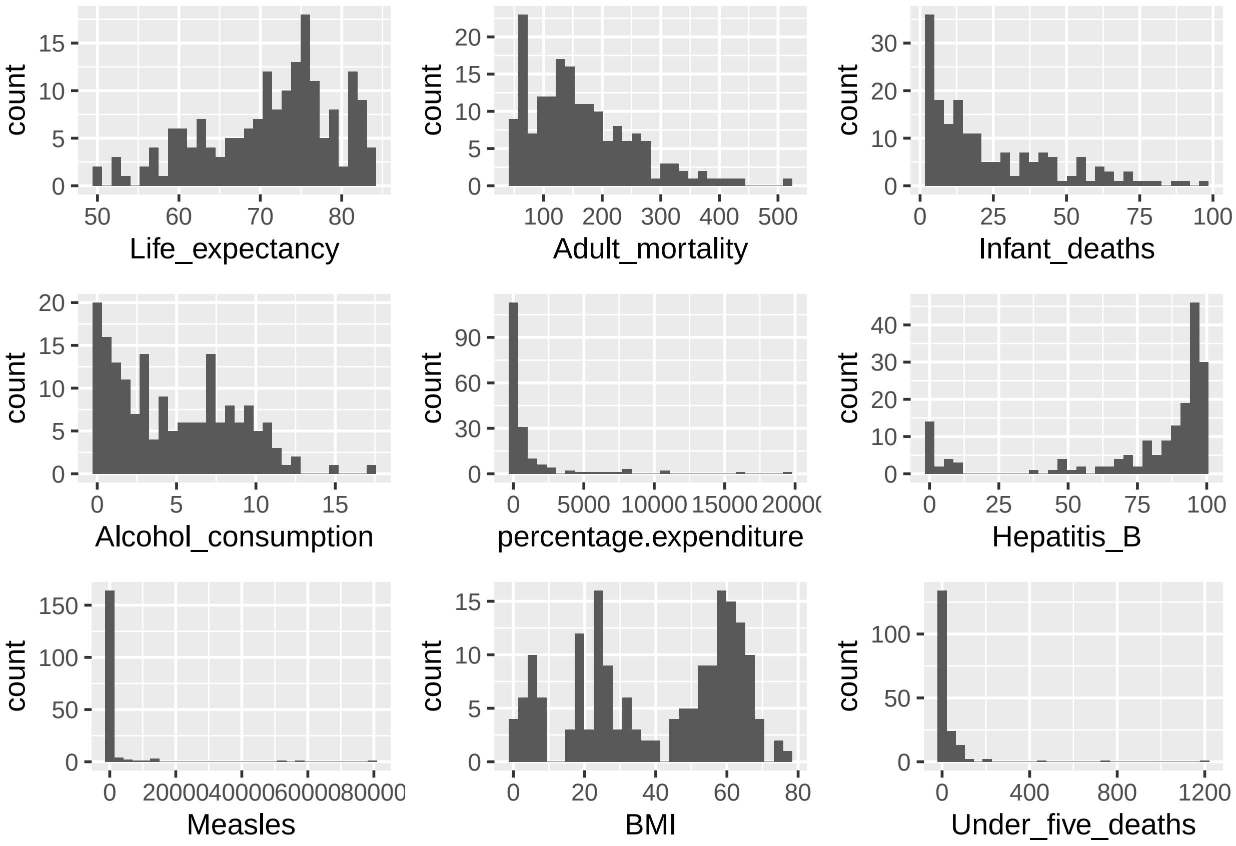
2.2 visualization

2.2.1 visualization: shape of distribution

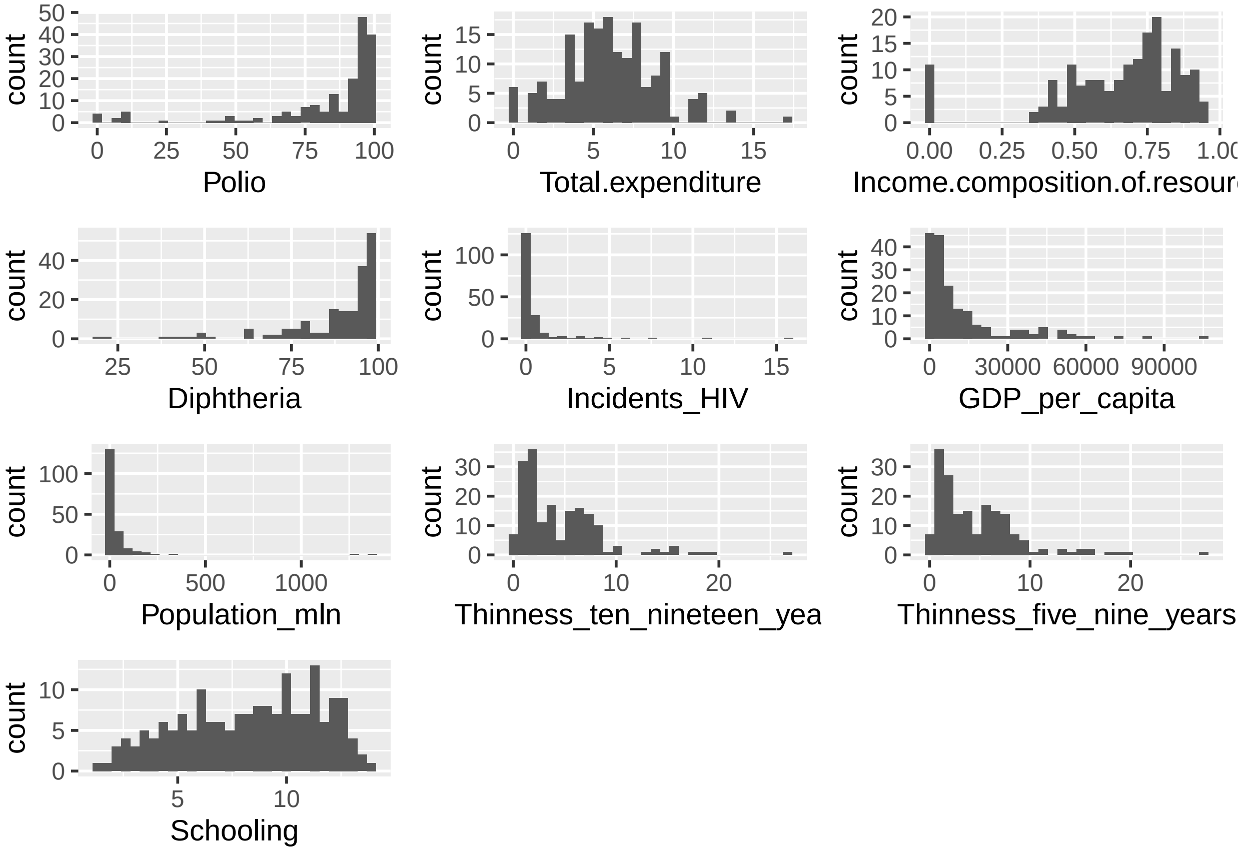
- Histogram for all the variables to visualize the frequency distribution of each. After plotting histograms, we dropped `Measles`, `Under_five_deaths`, `Incidents_HIV` due to a lot of zero values.

```
gg1 <- ggplot(data=led) + geom_histogram(aes(x=Life_expectancy), bins = 30)
gg2 <- ggplot(data=led) + geom_histogram(aes(x=Adult_mortality), bins = 30)
gg3 <- ggplot(data=led) + geom_histogram(aes(x=Infant_deaths), bins = 30)
gg4 <- ggplot(data=led) + geom_histogram(aes(x=Alcohol_consumption), bins = 30)
gg5 <- ggplot(data=led) + geom_histogram(aes(x=percentage.expenditure), bins = 30)
gg6 <- ggplot(data=led) + geom_histogram(aes(x=Hepatitis_B), bins = 30)
gg7 <- ggplot(data=led) + geom_histogram(aes(x=Measles), bins = 30)
gg8 <- ggplot(data=led) + geom_histogram(aes(x=BMI), bins = 30)
gg9 <- ggplot(data=led) + geom_histogram(aes(x=Under_five_deaths), bins = 30)
gg10 <- ggplot(data=led) + geom_histogram(aes(x=Polio), bins = 30)
gg11 <- ggplot(data=led) + geom_histogram(aes(x=Total.expenditure), bins = 30)
gg12 <- ggplot(data=led) + geom_histogram(aes(x=Income.composition.of.resources), bins = 30)
gg13 <- ggplot(data=led) + geom_histogram(aes(x=Diphtheria), bins = 30)
gg14 <- ggplot(data=led) + geom_histogram(aes(x=Incidents_HIV), bins = 30)
gg15 <- ggplot(data=led) + geom_histogram(aes(x=GDP_per_capita), bins = 30)
gg16 <- ggplot(data=led) + geom_histogram(aes(x=Population_mln), bins = 30)
gg17 <- ggplot(data=led) + geom_histogram(aes(x=Thinness_ten_nineteen_years), bins = 30)
```

```r
gg18 <- ggplot(data=led) + geom_histogram(aes(x=Thinness_five_nine_years), bins = 30)
gg19 <- ggplot(data=led) + geom_histogram(aes(x=Schooling), bins = 30)

grid.arrange(gg1, gg2, gg3, gg4, gg5, gg6, gg7, gg8, gg9, ncol=3)
```

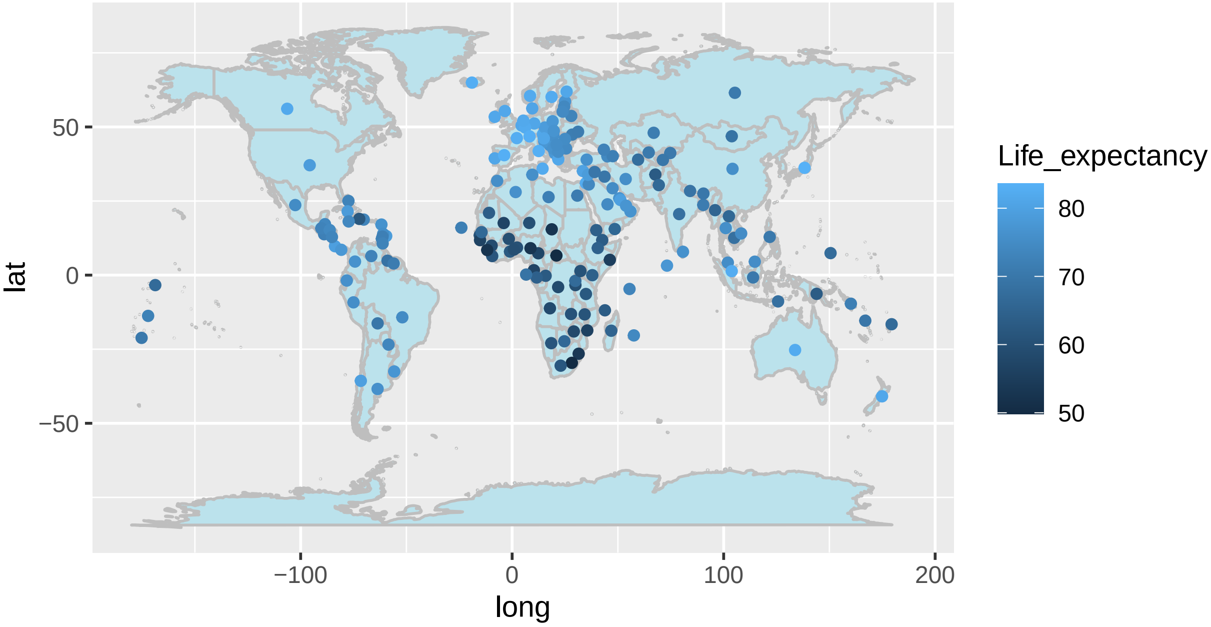


```r
grid.arrange(gg10, gg11, gg12, gg13, gg14,gg15, gg16, gg17, gg18, gg19, ncol=3)
```



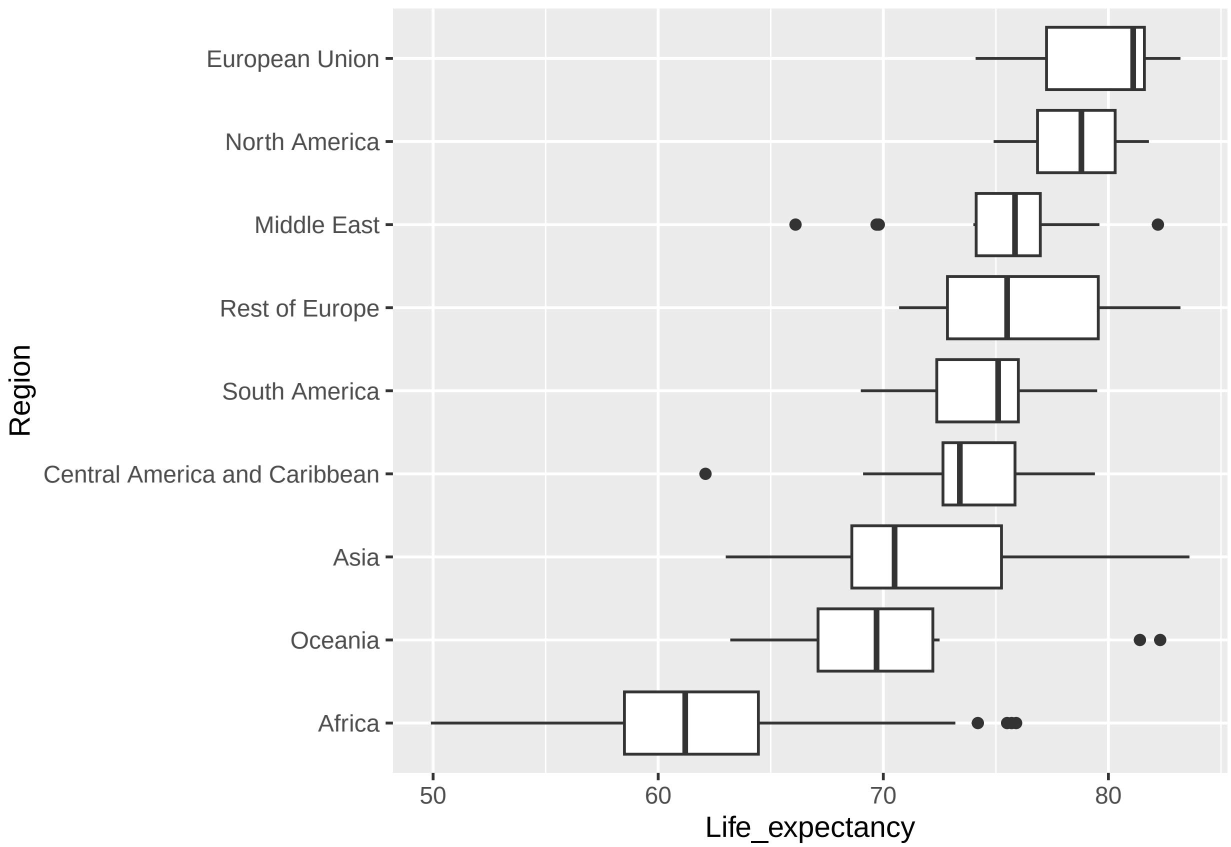
- Life expectancy by country on the world map : The size of each bubble reflects the magnitude of a certain variable, such as population or disease rates, at that specific location.The map highlights how the variables is distributed globally, with larger bubbles indicating higher values.

```
mapdata <- inner_join(led, countries, by=c("Country"="name"))

ggplot() +
  geom_polygon(data = map_data("world"), aes(x=long, y = lat, group = group)
               , fill="#BBE2EC", color="grey") +
  coord_fixed(1.4) +
  geom_point(data=mapdata, aes(x=longitude, y=latitude, color=Life_expectancy))
```



- `Life_expectancy`: Life expectancy by region. The boxplot displays life expectancy distributions across different global regions. The central box of each plot indicates the middle 50% of data, the line within represents the median, and the "whiskers" show the typical range. Regions like the European Union have high life expectancies, whereas Africa has a lower median and more variability, with outliers indicating extremes.

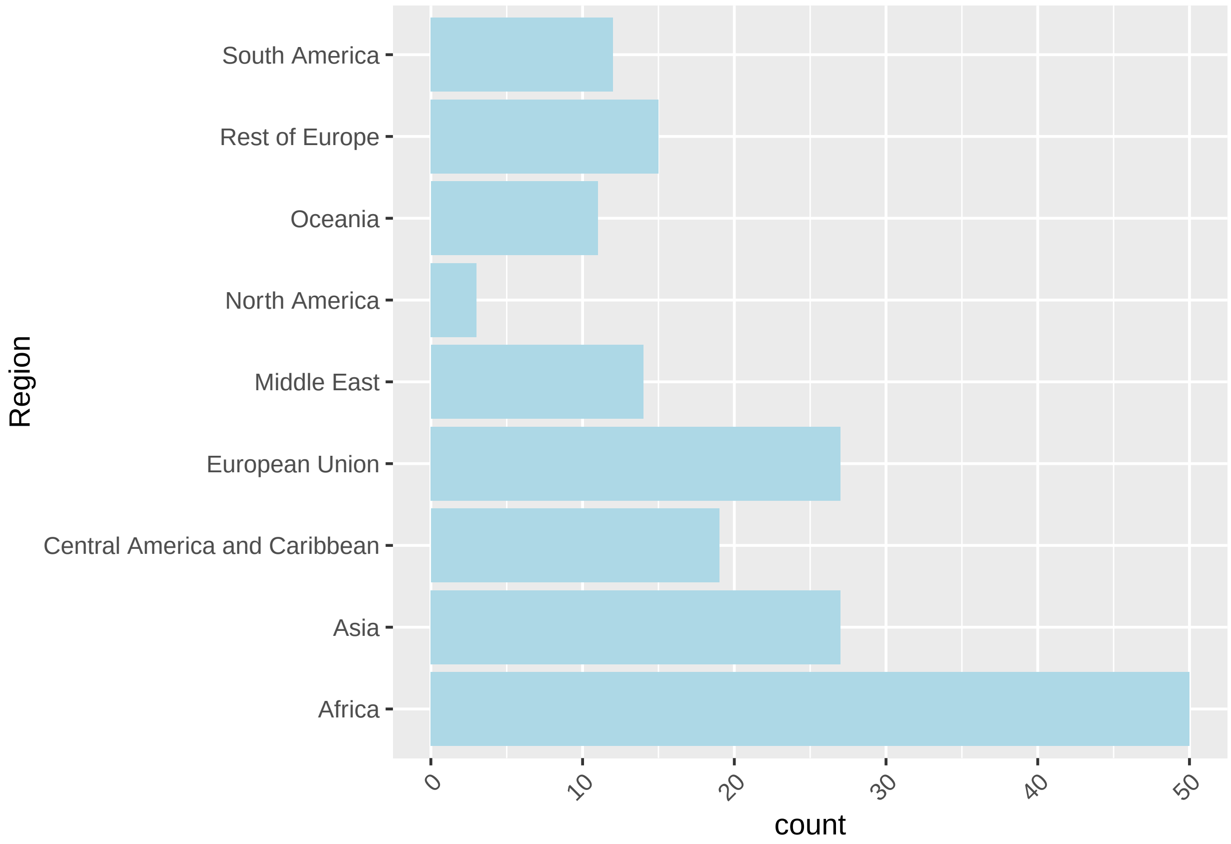
```r
ggplot(data=led) +
  geom_boxplot(aes(x=reorder(Region, Life_expectancy, FUN = median), y=Life_expectancy)) +
  xlab("Region") +
  coord_flip()
```



```
#filter(led, Life_expectancy <= 51)
```

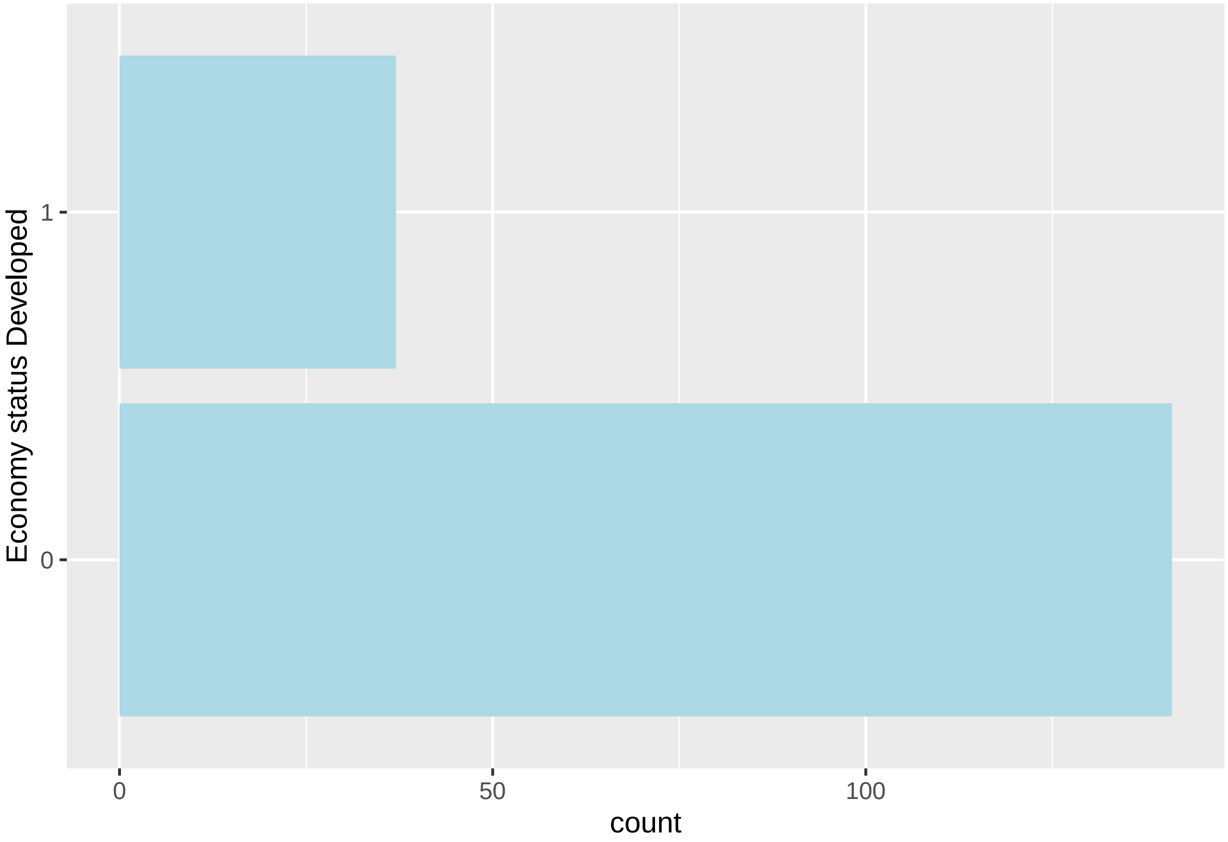
- `Region`: countries are distributed in 9 regions. The length of each bar represents the value of this variable for each region, with longer bars indicating higher values. The chart makes it easy to compare the metric across different regions, with Africa showing the highest value on the chart.

```
ggplot(data=led) +
  geom_bar(aes(x=Region), fill="lightblue") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  coord_flip()
```



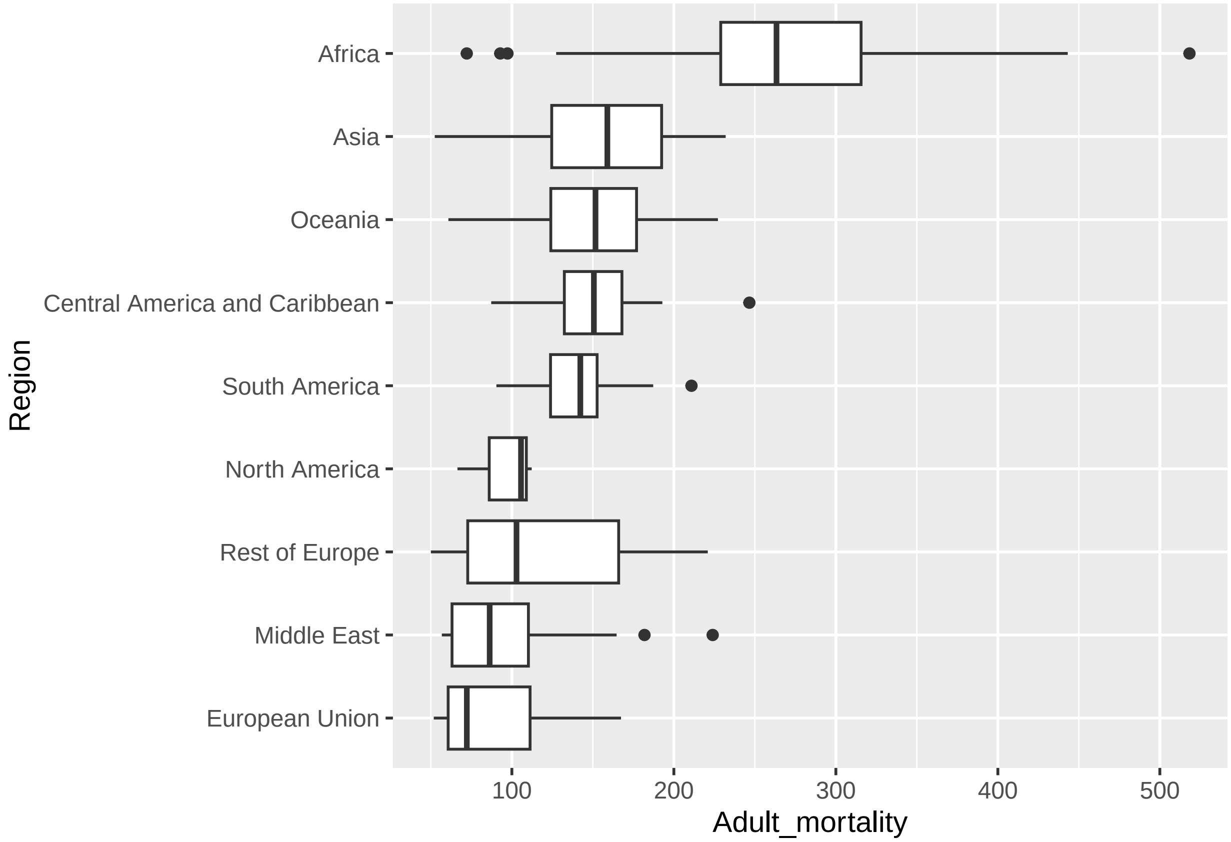
- `Economy_status_Developed`: 0 indicates it is a developing country while 1 indicates that it is a developed country

```
ggplot(data=led) +
  geom_bar(aes(x=as.factor(Economy_status_Developed)), fill="lightblue") +
  xlab("Economy status Developed") +
  coord_flip()
```



- Adult_mortality‘: Represents deaths of adults per 1000 population. The boxplot shows significant regional variations, with Africa showing a particularly high range and outliers indicating extreme mortality rates.

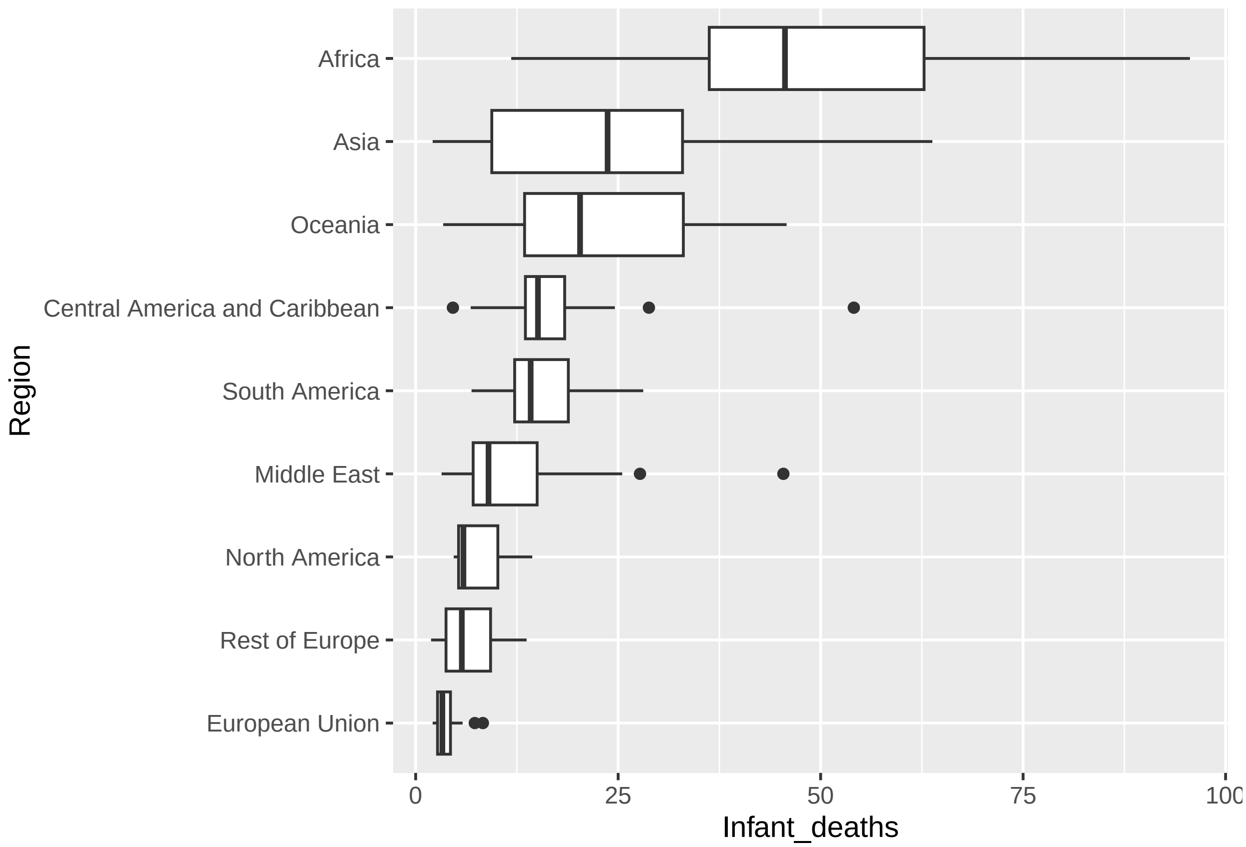
```
ggplot(data=led) +
  geom_boxplot(aes(x=reorder(Region, Adult_mortality, FUN = median), y=Adult_mortality)) +
  xlab("Region") +
  coord_flip()
```



```r
#filter(led, Adult_mortality > 350)
```

- `Infant_deaths`: Represents infant deaths per 1000 population. The boxplot layout allows easy comparison of infant mortality between these diverse regions. It shows median rates, the spread of the data, and outliers, with Africa displaying a notably wide range and higher rates of infant deaths compared to other regions.

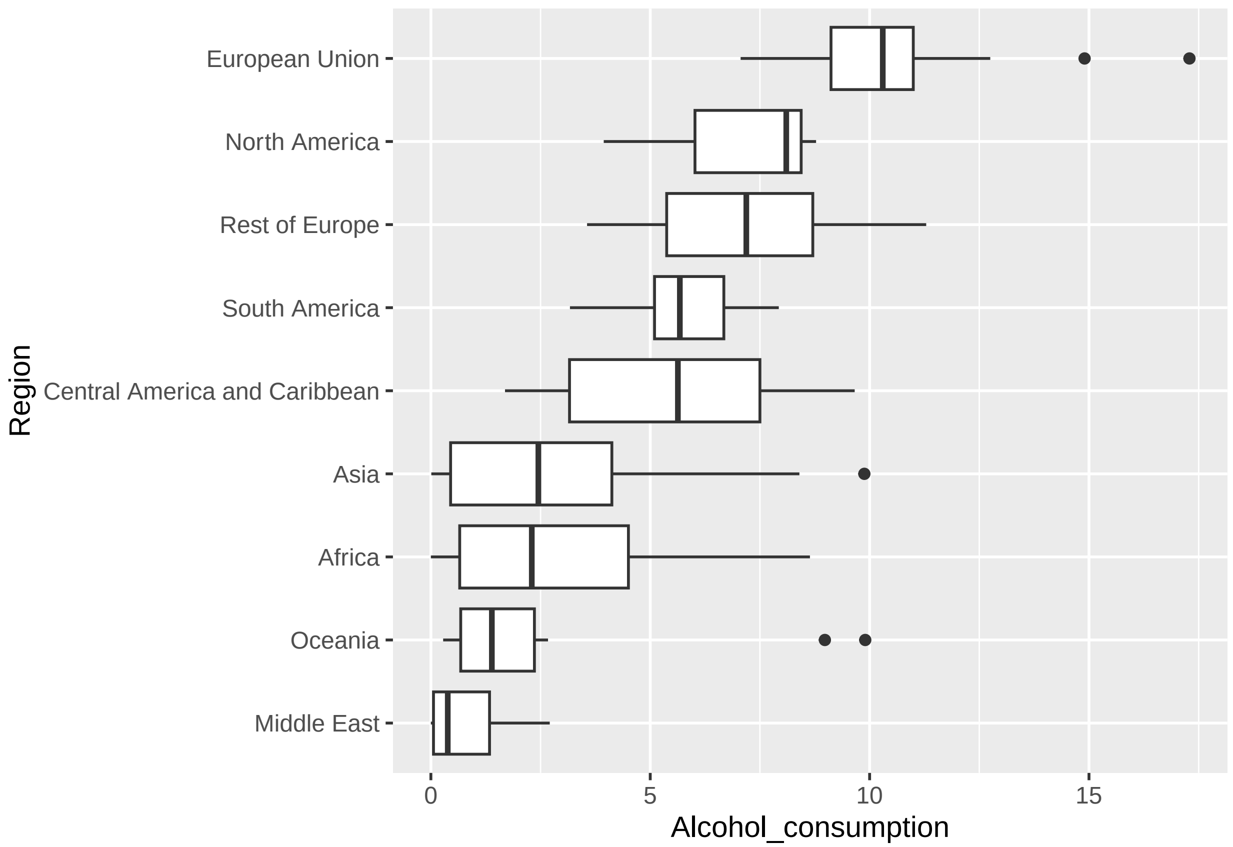
```r
ggplot(data=led) +
  geom_boxplot(aes(x=reorder(Region, Infant_deaths, FUN = median), y=Infant_deaths)) +
  xlab("Region") +
  coord_flip()
```



```
#filter(led, Adult_mortality > 350)
```

- `Alcohol_consumption`: The boxplot chart compares alcohol consumption, measured in liters of pure alcohol per capita for individuals aged 15 and older, across various global regions. The median consumption varies by region, with the European Union at the higher end and the Middle East at the lower. Outliers in several regions suggest significant variations from the median within those populations.

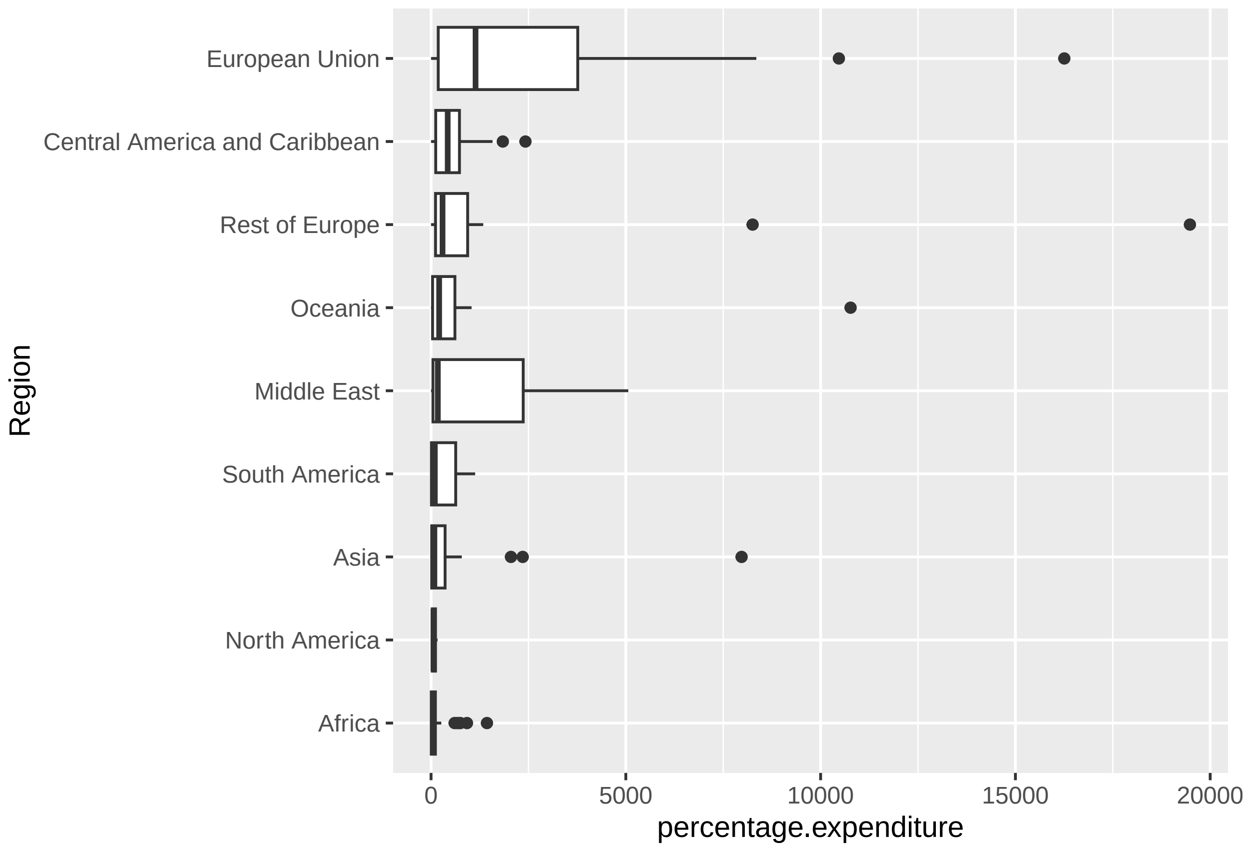
```
ggplot(data=led) +
  geom_boxplot(aes(x=reorder(Region, Alcohol_consumption, FUN = median), y=Alcohol_consumption)) +
  xlab("Region") +
  coord_flip()
```



```
#filter(led, Adult_mortality > 350)
```

- `percentage.expenditure`: The boxplot chart that displays health expenditure as a percentage of Gross Domestic Product (GDP) per capita across different regions. The European Union has the widest range of expenditures, indicating significant variation within the region. Africa, while showing lower median expenditure, has outliers suggesting a few cases of relatively high spending.

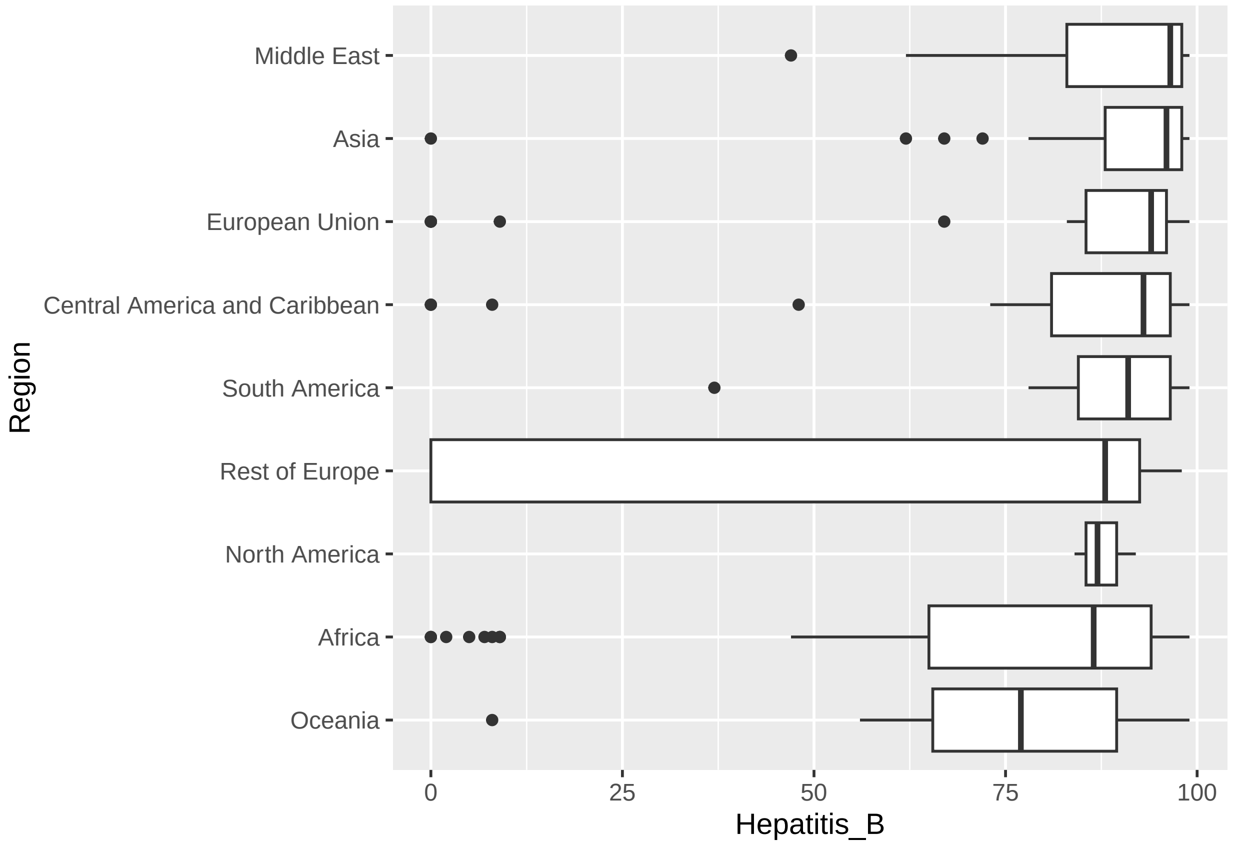
```
ggplot(data=led) +
  geom_boxplot(aes(x=reorder(Region, percentage.expenditure, FUN = median)
                   , y=percentage.expenditure)) +
  xlab("Region") +
  coord_flip()
```



```r
#filter(led, Adult_mortality > 350)
```

- $Hepatitis_B: The boxplot chart showcases the percentage coverage of Hepatitis B immunization among 1-year-olds in various regions. North America and the European Union show high median coverage rates with relatively small interquartile ranges, while Africa exhibits lower coverage rates with greater variability. Some regions have outliers, indicating substantial deviations from the median coverage rates.

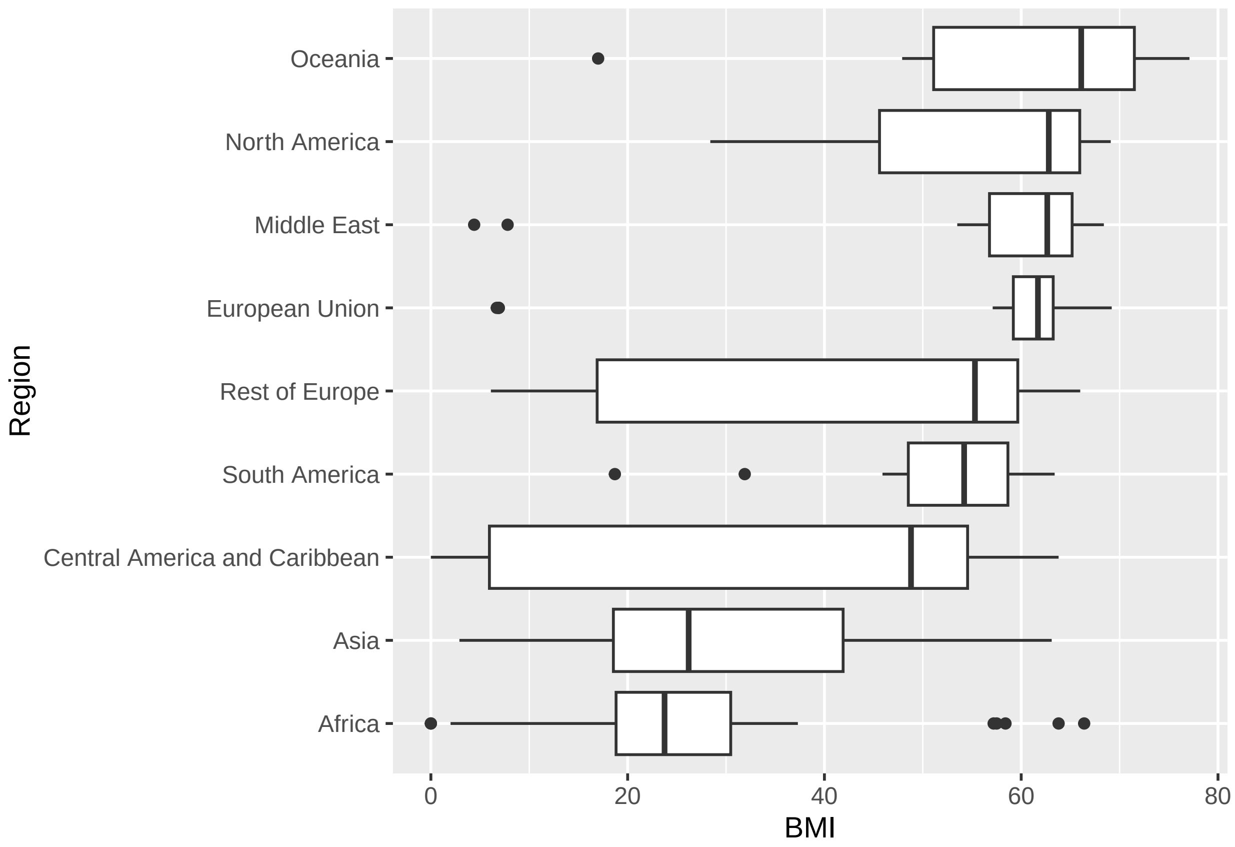
```r
ggplot(data=led) +
  geom_boxplot(aes(x=reorder(Region, Hepatitis_B, FUN = median), y=Hepatitis_B)) +
  xlab("Region") +
  coord_flip()
```



```
#filter(led, Adult_mortality > 350)
```

- BMI': The boxplot illustrates the average Body Mass Index (BMI) across various regions. It highlights the median BMI, the spread of BMI values within each region, and outliers. The regions show varying levels of BMI, with some like Oceania and North America exhibiting higher median BMIs, and others like Africa having lower median values with outliers representing extremely low BMIs.

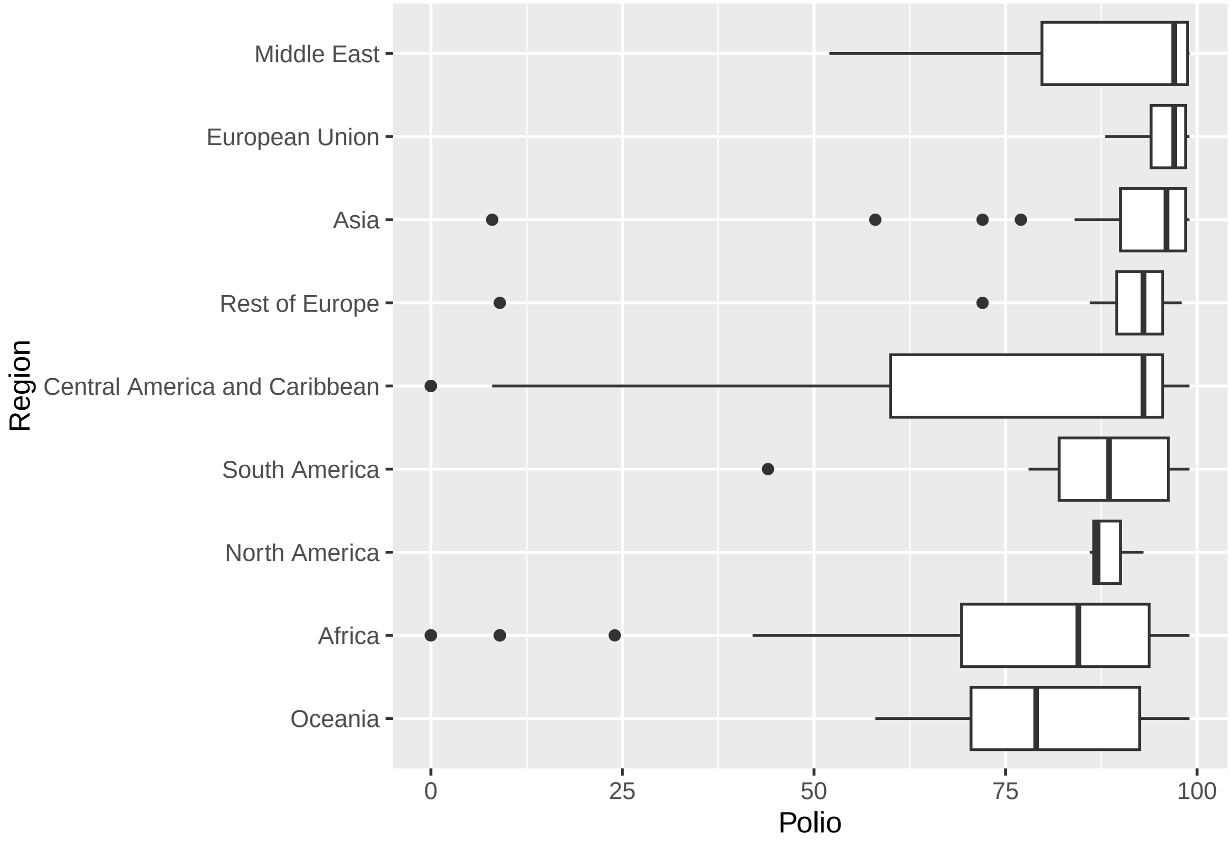
```
ggplot(data=led) +
  geom_boxplot(aes(x=reorder(Region, BMI, FUN = median), y=BMI)) +
  xlab("Region") +
  coord_flip()
```



```
#filter(led, Adult_mortality > 350)
```

- `Polio`: The boxplot displays the percentage of 1-year-olds who have received the polio vaccine across various regions. Most regions have high median immunization coverage with some showing more variability than others. Africa and Asia have some outliers indicating areas with lower than typical immunization rates.

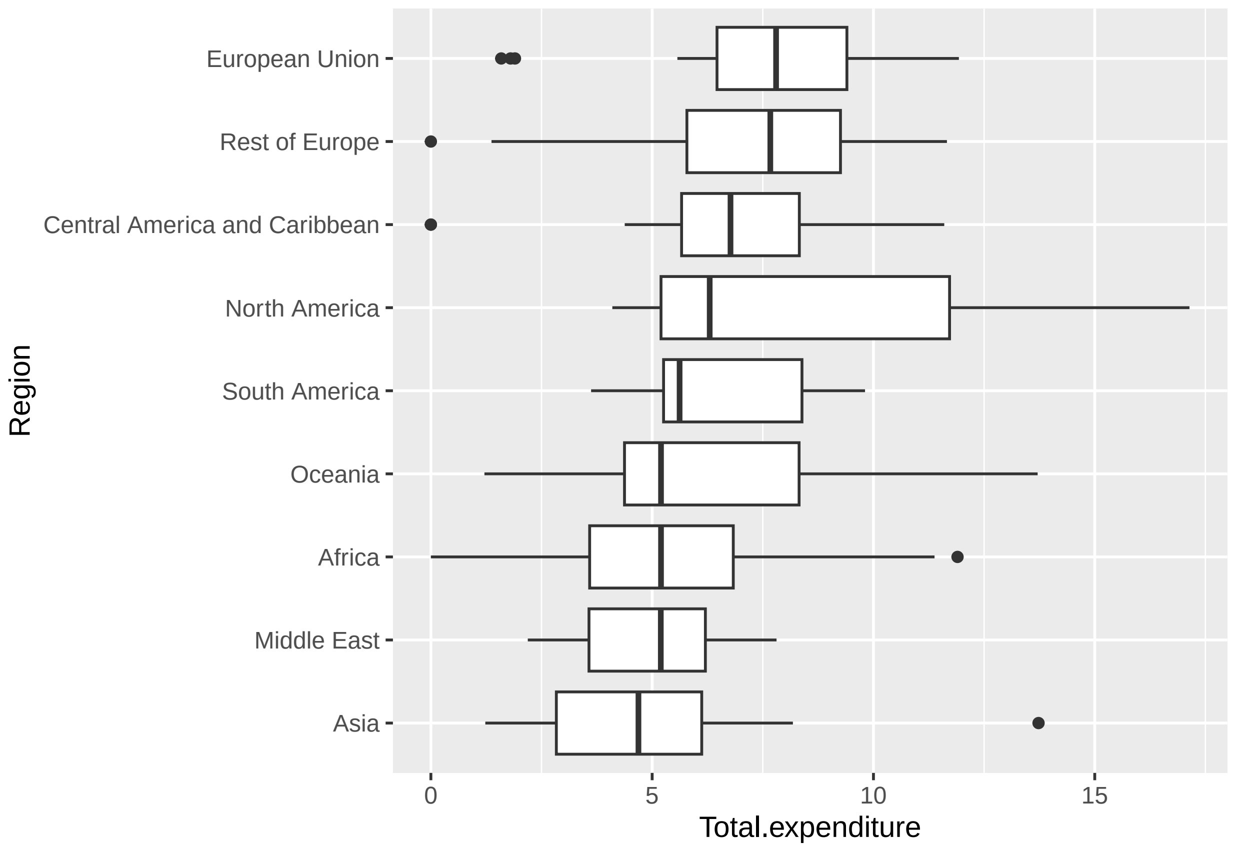
```
ggplot(data=led) +
  geom_boxplot(aes(x=reorder(Region, Polio, FUN = median), y=Polio)) +
  xlab("Region") +
  coord_flip()
```



```
#filter(led, Adult_mortality > 350)
```

- `Total.expenditure`:The boxplot depicts the percentage of total government expenditure dedicated to health across various regions. It shows a wide range of health spending percentages, with some regions like the European Union displaying higher median expenditures and significant outliers indicating both higher and lower spending levels. Regions like Asia and the Middle East show lower median government health expenditures with fewer outliers.

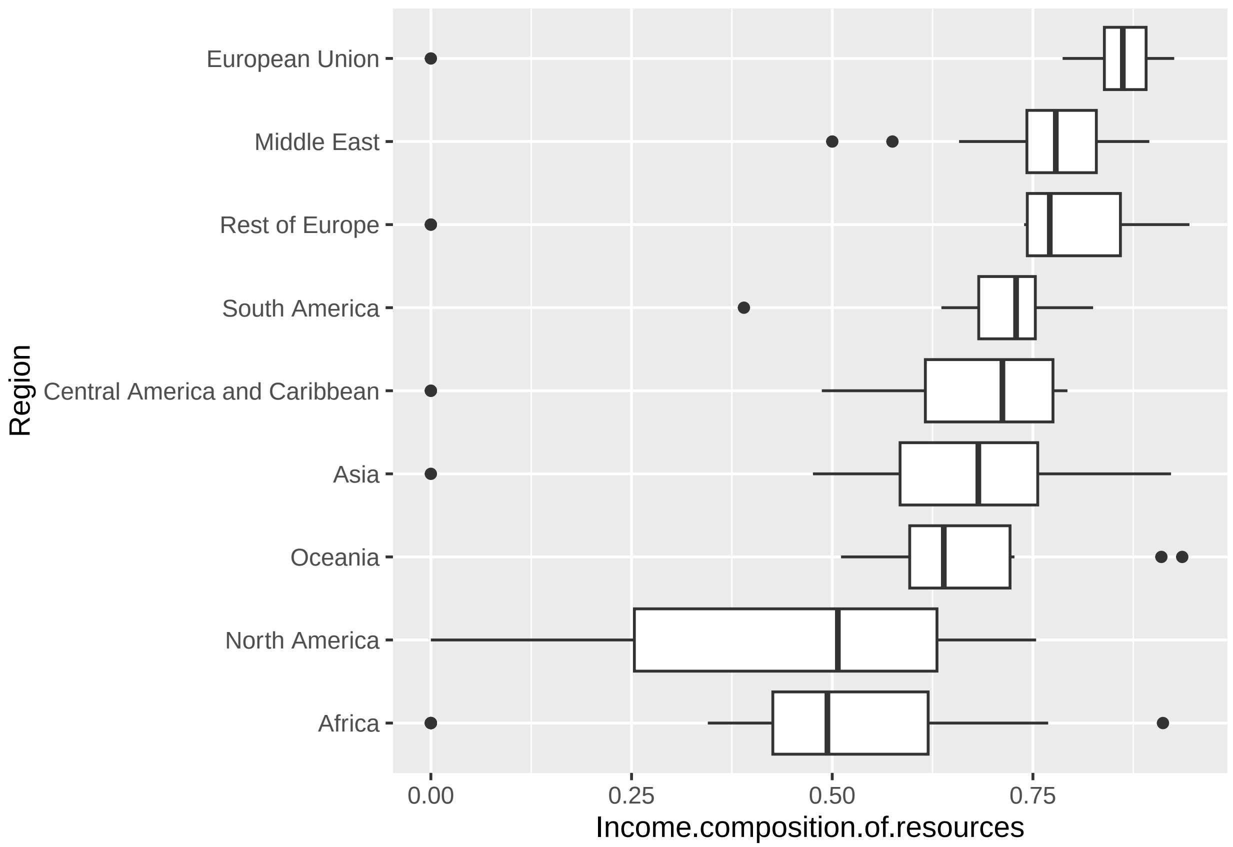
```
ggplot(data=led) +
  geom_boxplot(aes(x=reorder(Region, Total.expenditure, FUN = median)
                   , y=Total.expenditure)) +
  xlab("Region") +
  coord_flip()
```



```
#filter(led, Adult_mortality > 350)
```

- `Income.composition.of.resources`: The boxplot that visualizes the distribution of DTP3 immunization coverage among 1-year-olds across various regions. Regions like North America and the European Union exhibit high median immunization coverage with relatively little variability, while Africa shows lower median coverage and greater variability, with some outliers indicating very low coverage rates.

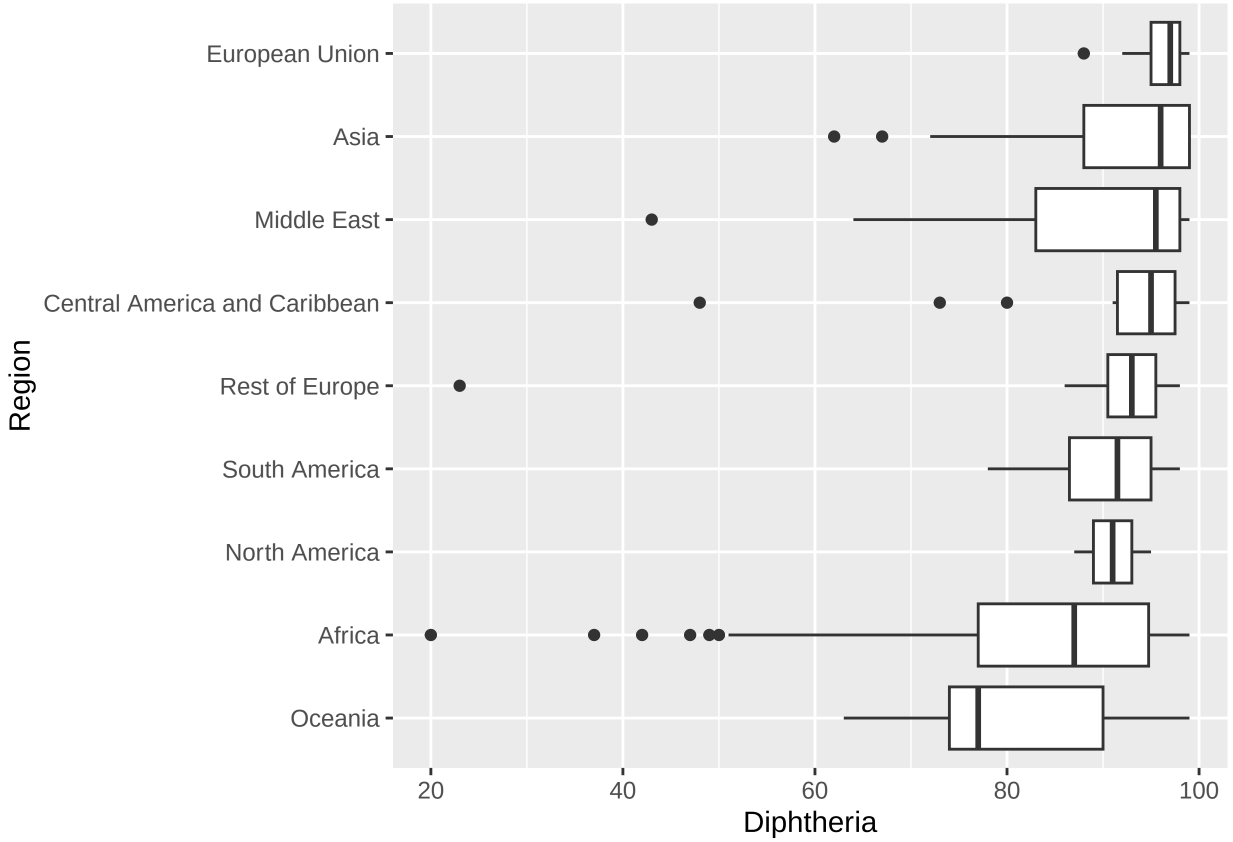
```
ggplot(data=led) +
  geom_boxplot(aes(x=reorder(Region, Income.composition.of.resources, FUN = median), y=Income.compositi
  xlab("Region") +
  coord_flip()
```



```
#filter(led, Adult_mortality > 350)
```

- `Diphtheria`: The boxplot illustrates the incidence of HIV per 1000 population aged 15-49 across various regions. The European Union and North America have higher median incidences with wide interquartile ranges, indicating variability, while Africa displays both a higher median and several outliers, reflecting extremely high incidences in some areas.

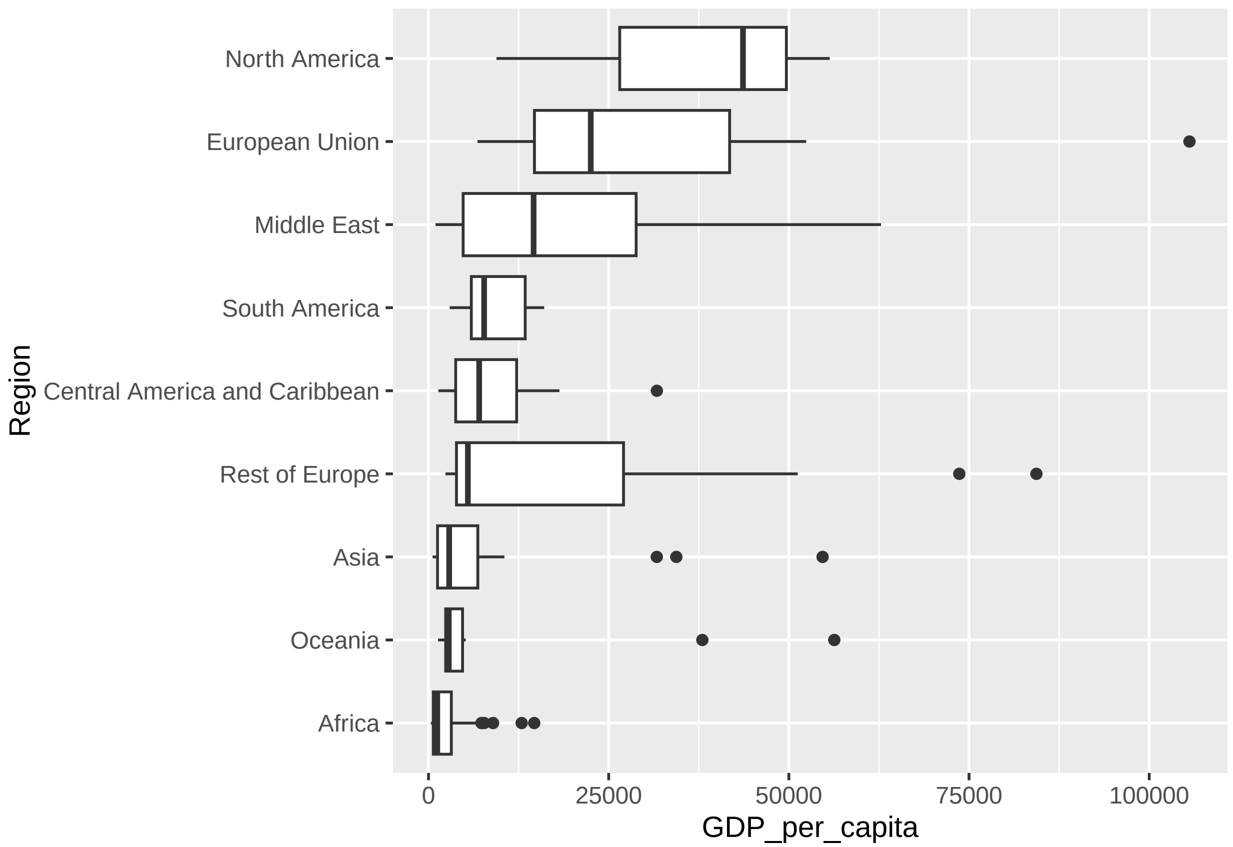
```
ggplot(data=led) +
  geom_boxplot(aes(x=reorder(Region, Diphtheria, FUN = median), y=Diphtheria)) +
  xlab("Region") +
  coord_flip()
```



```
#filter(led, Adult_mortality > 350)
```

- `GDP_per_capita`: The boxplot visualizes the GDP per capita in USD across various regions, with each box indicating the spread of GDP values within that region. North America and the European Union display higher medians and wider ranges, indicating more variability, while Africa shows the lowest median GDP per capita, with several outliers indicating instances of extremely low GDP.

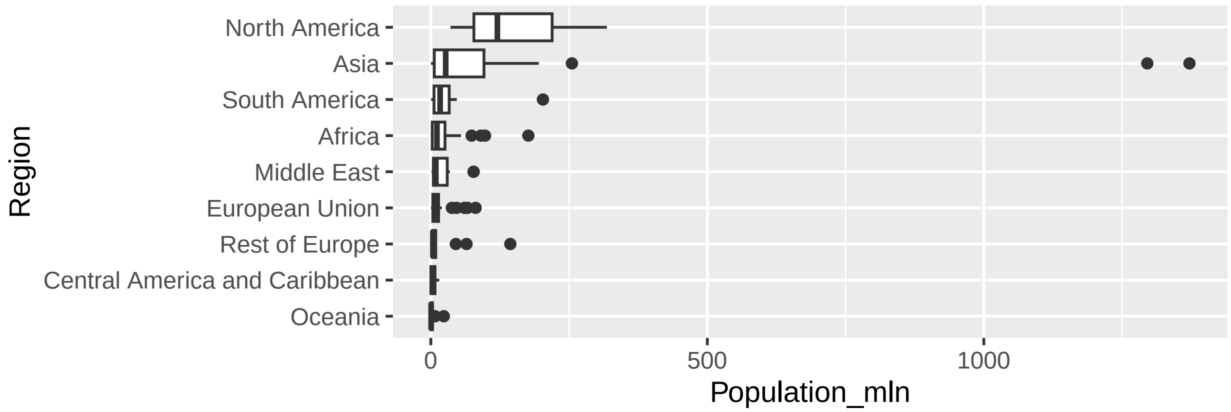
```
ggplot(data=led) +
  geom_boxplot(aes(x=reorder(Region, GDP_per_capita, FUN = median), y=GDP_per_capita)) +
  xlab("Region") +
  coord_flip()
```

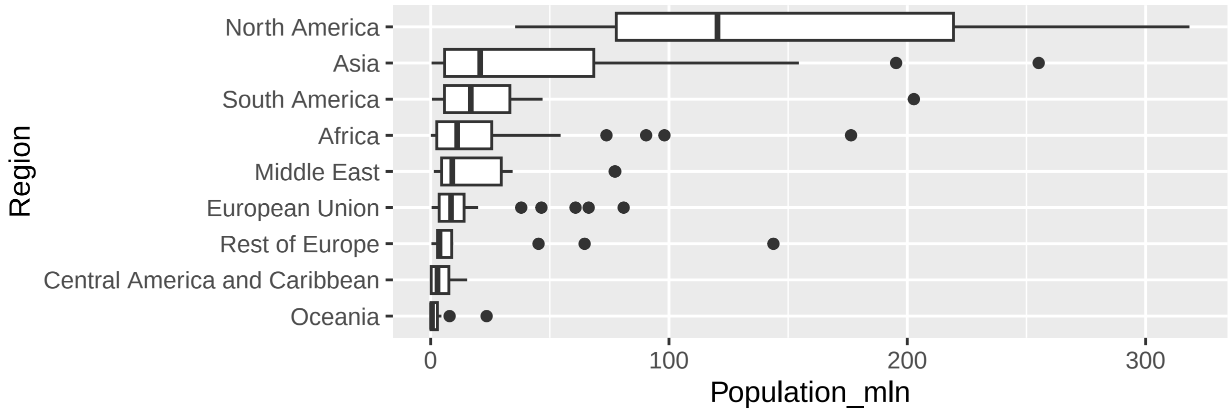


```
#filter(led, Adult_mortality > 350)
```

- `Population_mln`: The boxplots depicts total population in millions for various global regions. The top plot has a wider scale, showing regions with large populations like Asia, while the bottom plot has a narrower scale, providing a more detailed look at regions with smaller populations. Both illustrate the spread and central tendency of regional populations, with outliers representing significantly smaller or larger populations compared to the median of the regions.

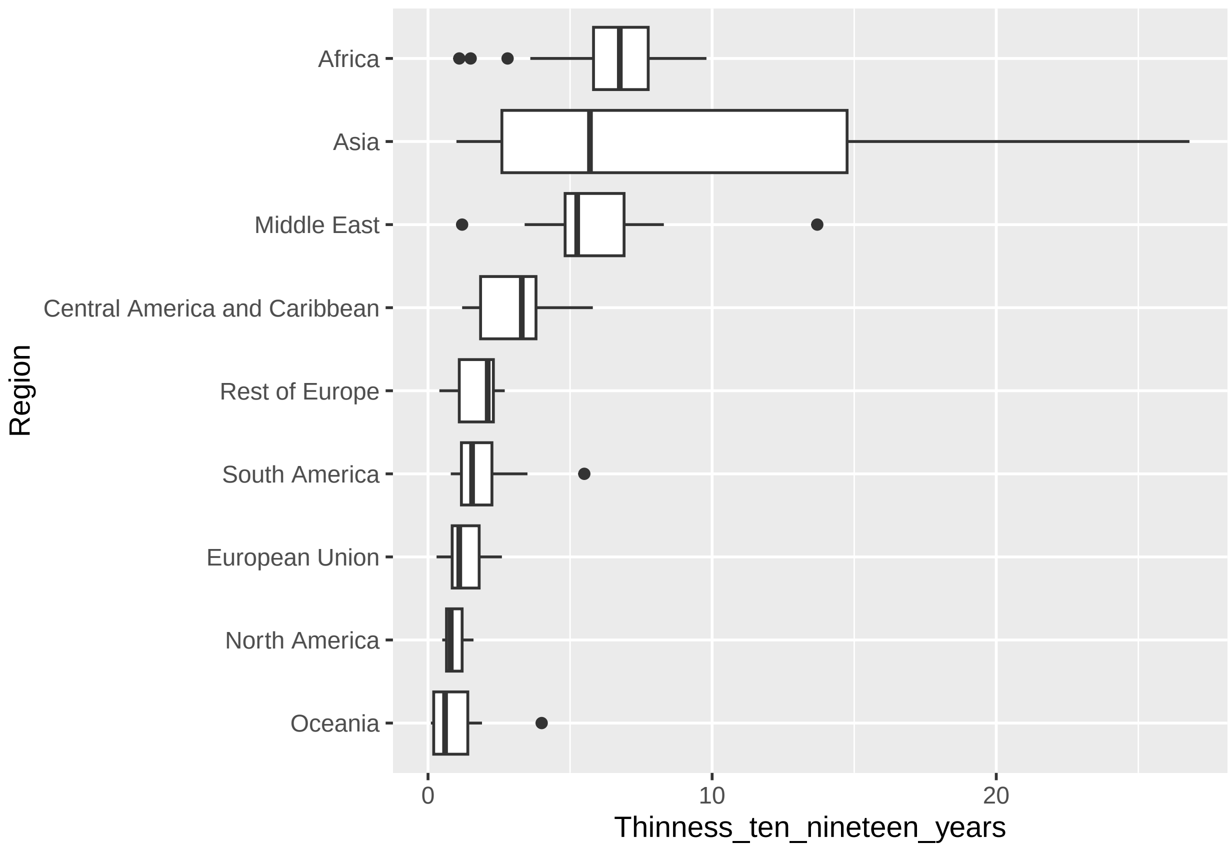
```
gg1 <- ggplot(data=led) +
  geom_boxplot(aes(x=reorder(Region, Population_mln, FUN = median), y=Population_mln)) +
  xlab("Region") +
  coord_flip()

#filter(led, Population_mln > 1000)

led_pop <- filter(led, Population_mln < 1000)

gg2 <- ggplot(data=led_pop) +
  geom_boxplot(aes(x=reorder(Region, Population_mln, FUN = median), y=Population_mln)) +
  xlab("Region") +
  coord_flip()

grid.arrange(gg1, gg2, nrow=2)
```





- `Thinness_ten_nineteen_years`:The boxplot that compares the prevalence of thinness among adolescents aged 10-19 across various regions. Africa shows a higher median prevalence and a wide interquartile range, indicating variability within the region, while other regions generally exhibit lower levels of adolescent thinness. Some regions also show outliers, reflecting extreme cases of thinness.

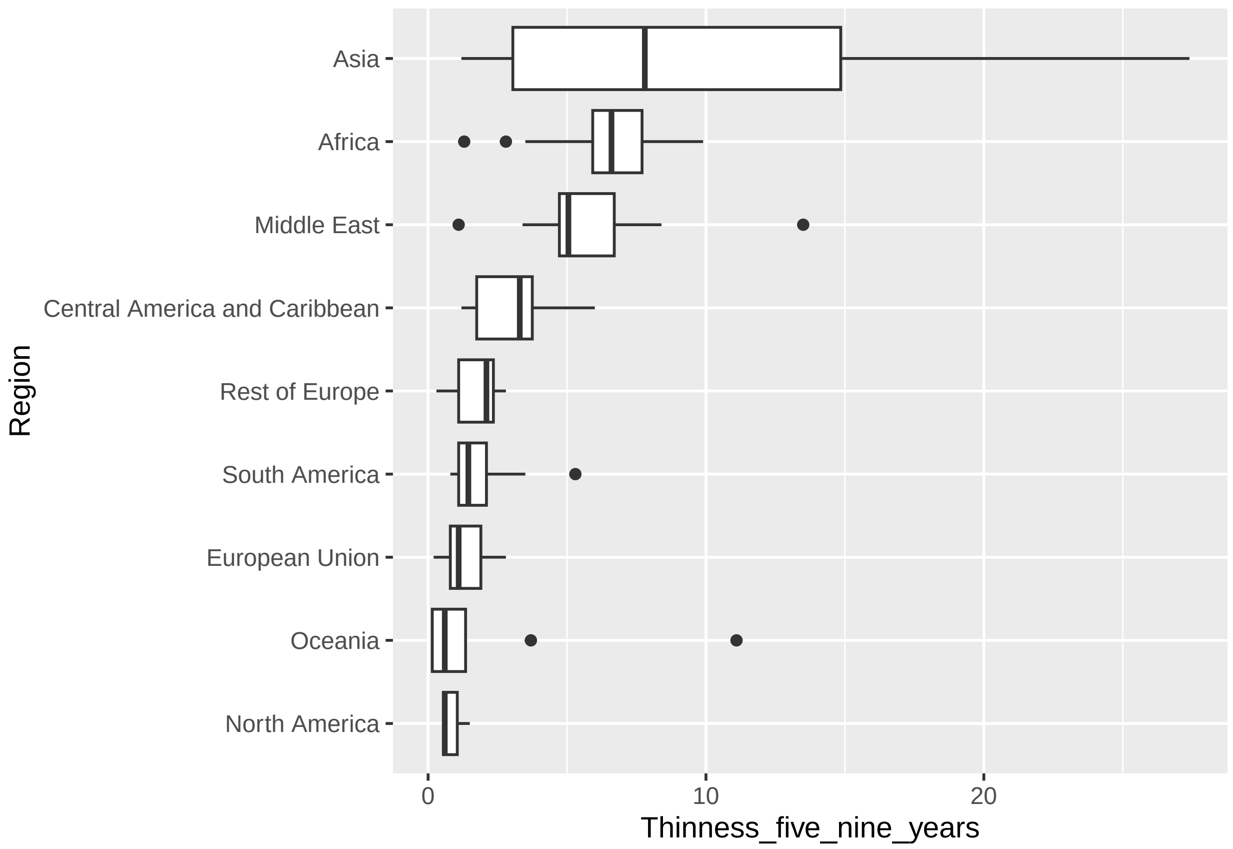
```
ggplot(data=led) +
  geom_boxplot(aes(x=reorder(Region, Thinness_ten_nineteen_years, FUN = median), y=Thinness_ten_ninetee
  xlab("Region") +
  coord_flip()
```



```
#filter(led, Adult_mortality > 350)
```

- `Thinness_five_nine_years` : The boxplot shows the prevalence of thinness among children aged 5-9 years across various regions, where thinness is defined as a BMI less than -2 standard deviations below the median. Africa and Asia show a higher prevalence and wider spread of thinness, while North America and the European Union have lower medians indicating a smaller prevalence of thinness in these regions. There are also several outliers, particularly in regions like Africa, highlighting greater variability within those populations.

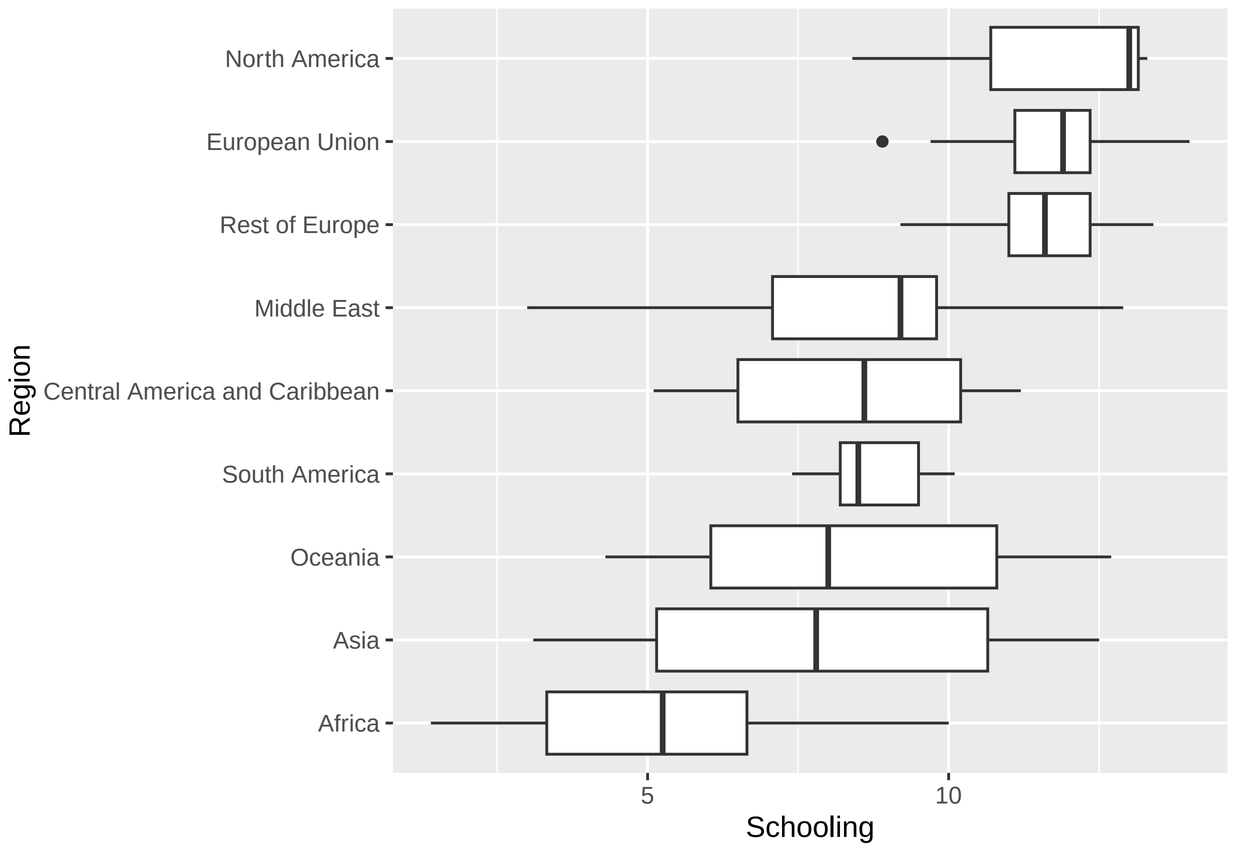
```
ggplot(data=led) +
  geom_boxplot(aes(x=reorder(Region, Thinness_five_nine_years, FUN = median)
                   , y=Thinness_five_nine_years)) +
  xlab("Region") +
  coord_flip()
```



```
#filter(led, Adult_mortality > 350)
```

- `Schooling`: The boxplot displays the average years of schooling for people aged 25 and older across various global regions. The data show significant regional differences, with North America and the European Union having higher averages and less variability, while Africa has the lowest average years of schooling and greater variability, as indicated by the length of the box and the position of the median.

```
ggplot(data=led) +
  geom_boxplot(aes(x=reorder(Region, Schooling, FUN = median), y=Schooling)) +
  xlab("Region") +
  coord_flip()
```



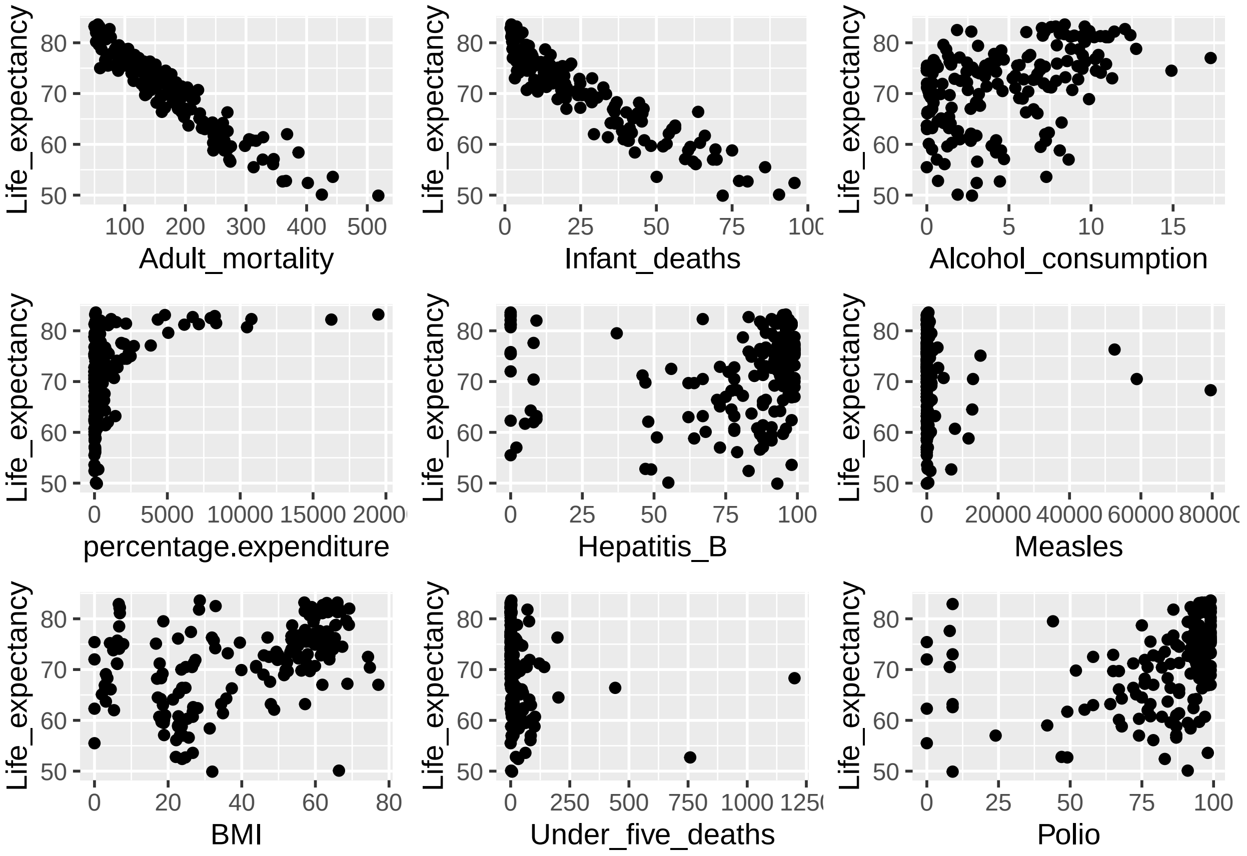
```
#filter(led, Adult_mortality > 350)
```

2.2.2 visualization: relationship between variables

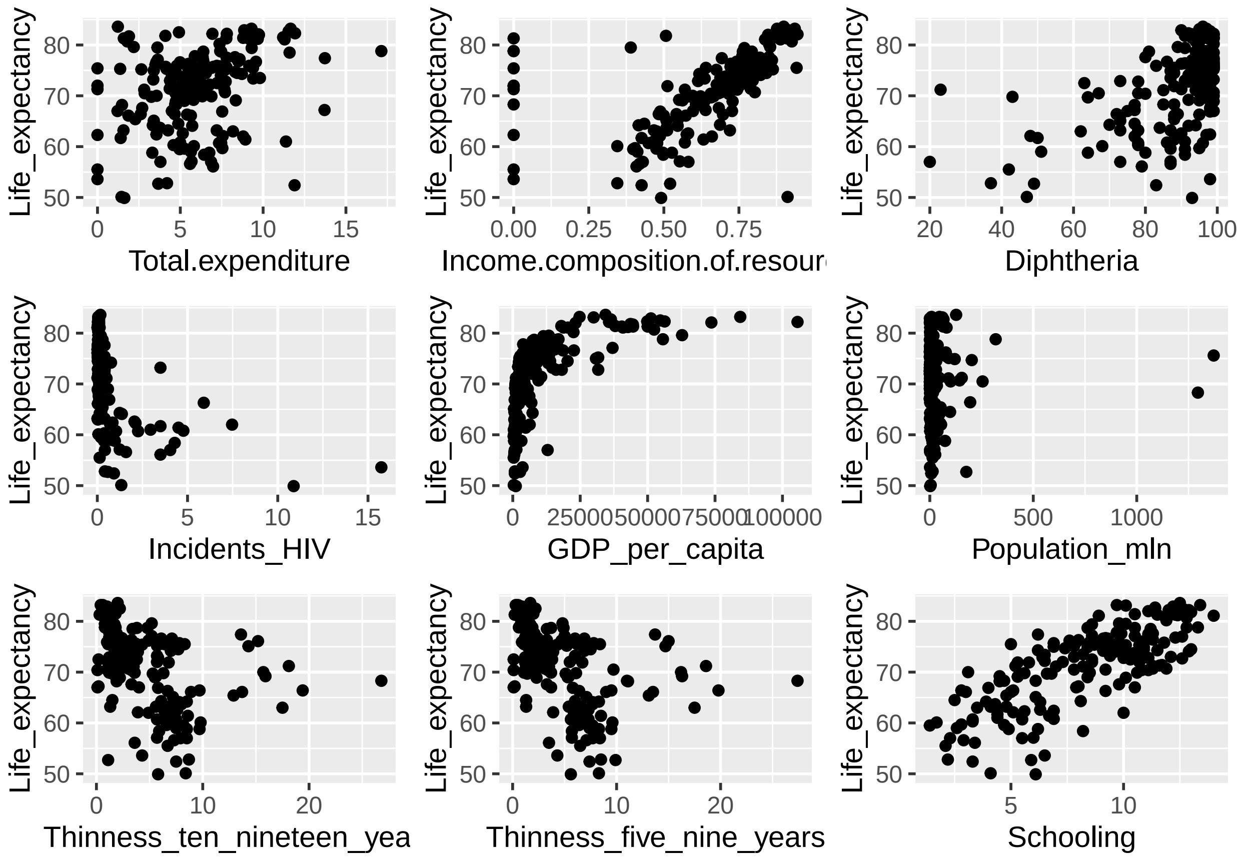
- Scatterplots : The scatterplots correlate life expectancy with various factors: higher adult mortality, infant and under-five deaths, and BMI generally associate with lower life expectancy, while increased Hepatitis B and polio vaccinations align with higher life expectancy. Alcohol consumption and measles cases display no clear pattern, and the impact of health expenditure on life expectancy is ambiguous, with a cluster of high life expectancy at lower expenditure levels. The scatterplots also show that higher life expectancy is generally associated with higher income, education levels, and immunization coverage, while negative correlations are observed with higher HIV prevalence and thinness among children and adolescents. Total health expenditure and population size do not exhibit a clear relationship with life expectancy.

```
gg1 <- ggplot(data=led) + geom_point(aes(x=Adult_mortality, y=Life_expectancy))
gg2 <- ggplot(data=led) + geom_point(aes(x=Infant_deaths, y=Life_expectancy))
gg3 <- ggplot(data=led) + geom_point(aes(x=Alcohol_consumption, y=Life_expectancy))
gg4 <- ggplot(data=led) + geom_point(aes(x=percentage.expenditure, y=Life_expectancy))
gg5 <- ggplot(data=led) + geom_point(aes(x=Hepatitis_B, y=Life_expectancy))
gg6 <- ggplot(data=led) + geom_point(aes(x=Measles, y=Life_expectancy))
gg7 <- ggplot(data=led) + geom_point(aes(x=BMI, y=Life_expectancy))
gg8 <- ggplot(data=led) + geom_point(aes(x=Under_five_deaths, y=Life_expectancy))
gg9 <- ggplot(data=led) + geom_point(aes(x=Polio, y=Life_expectancy))
gg10 <- ggplot(data=led) + geom_point(aes(x=Total.expenditure, y=Life_expectancy))
gg11 <- ggplot(data=led) + geom_point(aes(x=Income.composition.of.resources, y=Life_expectancy))
```

```
gg12 <- ggplot(data=led) + geom_point(aes(x=Diphtheria, y=Life_expectancy))
gg13 <- ggplot(data=led) + geom_point(aes(x=Incidents_HIV, y=Life_expectancy))
gg14 <- ggplot(data=led) + geom_point(aes(x=GDP_per_capita, y=Life_expectancy))
gg15 <- ggplot(data=led) + geom_point(aes(x=Population_mln, y=Life_expectancy))
gg16 <- ggplot(data=led) + geom_point(aes(x=Thinness_ten_nineteen_years, y=Life_expectancy))
gg17 <- ggplot(data=led) + geom_point(aes(x=Thinness_five_nine_years, y=Life_expectancy))
gg18 <- ggplot(data=led) + geom_point(aes(x=Schooling, y=Life_expectancy))

grid.arrange(gg1, gg2, gg3, gg4, gg5, gg6, gg7, gg8, gg9, ncol=3)
```

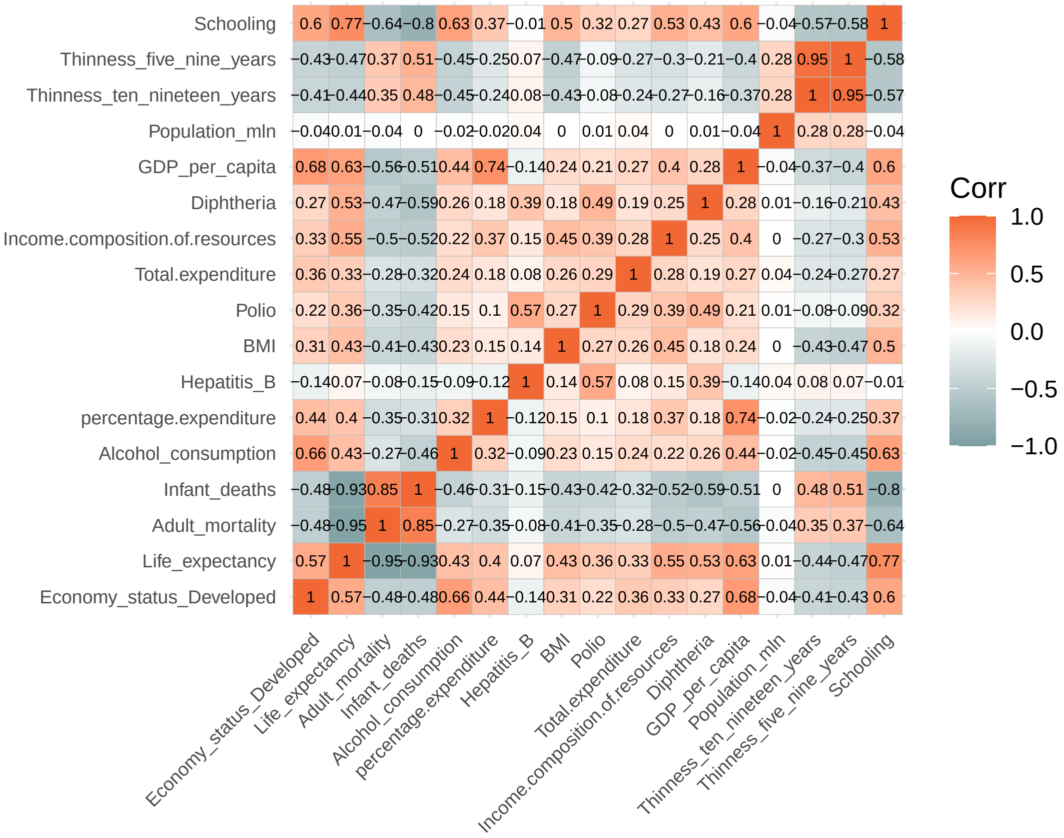


```
grid.arrange(gg10, gg11, gg12, gg13, gg14, gg15, gg16, gg17, gg18, ncol=3)
```



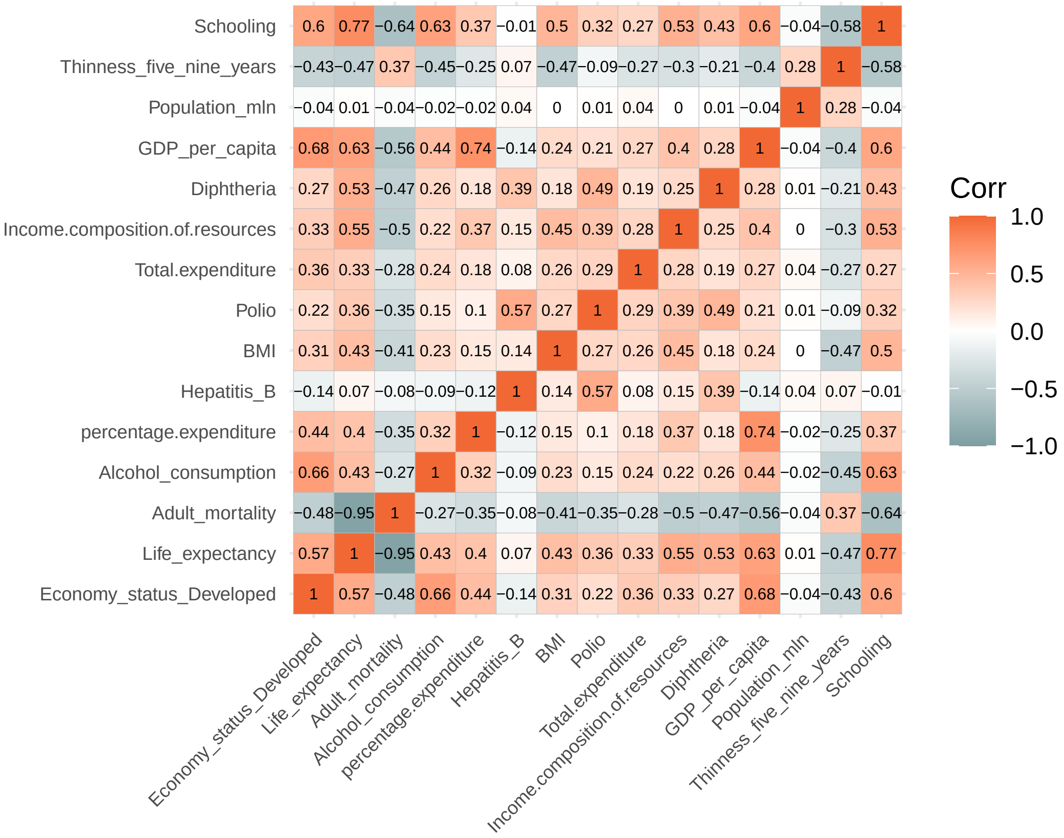
- Correlation Matrix

```
cor_led <- cor(select(led, c(-Country,-Region, -Incidents_HIV, -Under_five_deaths, -Measles)))
ggcorrplot(cor_led, lab = TRUE, lab_size=2, tl.cex = 7,
           sig.level = 0.05, digits = 2,
           colors = c("#7d9ea2", "white", "#f26835"))
```



- We drop `Infant_deaths` due to high correlation value 0.85 with `Adult_mortality`.
- We drop `Thinness_ten_nineteen_years` due to high correlation value 0.95 with `Thinness_five_nine_years`

```
cor_led <- cor(select(led, c(-Country,-Region, -Incidents_HIV, -Under_five_deaths, -Measles
                             , -Infant_deaths, -Thinness_ten_nineteen_years)))
ggcorrplot(cor_led, lab = TRUE, lab_size=2, tl.cex = 7,
           sig.level = 0.05, digits = 2,
           colors = c("#7d9ea2", "white", "#f26835"))
```

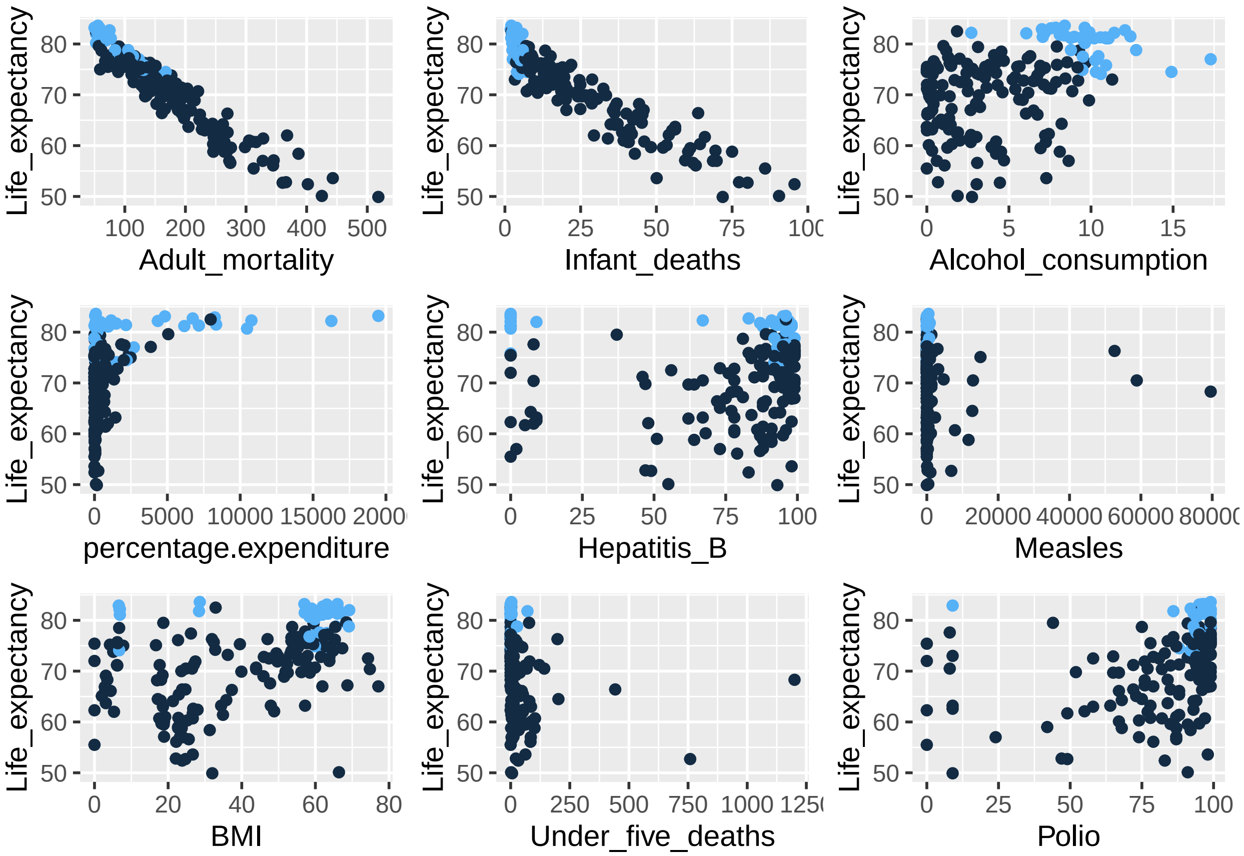


2.2.3 visualization: comparison across groups

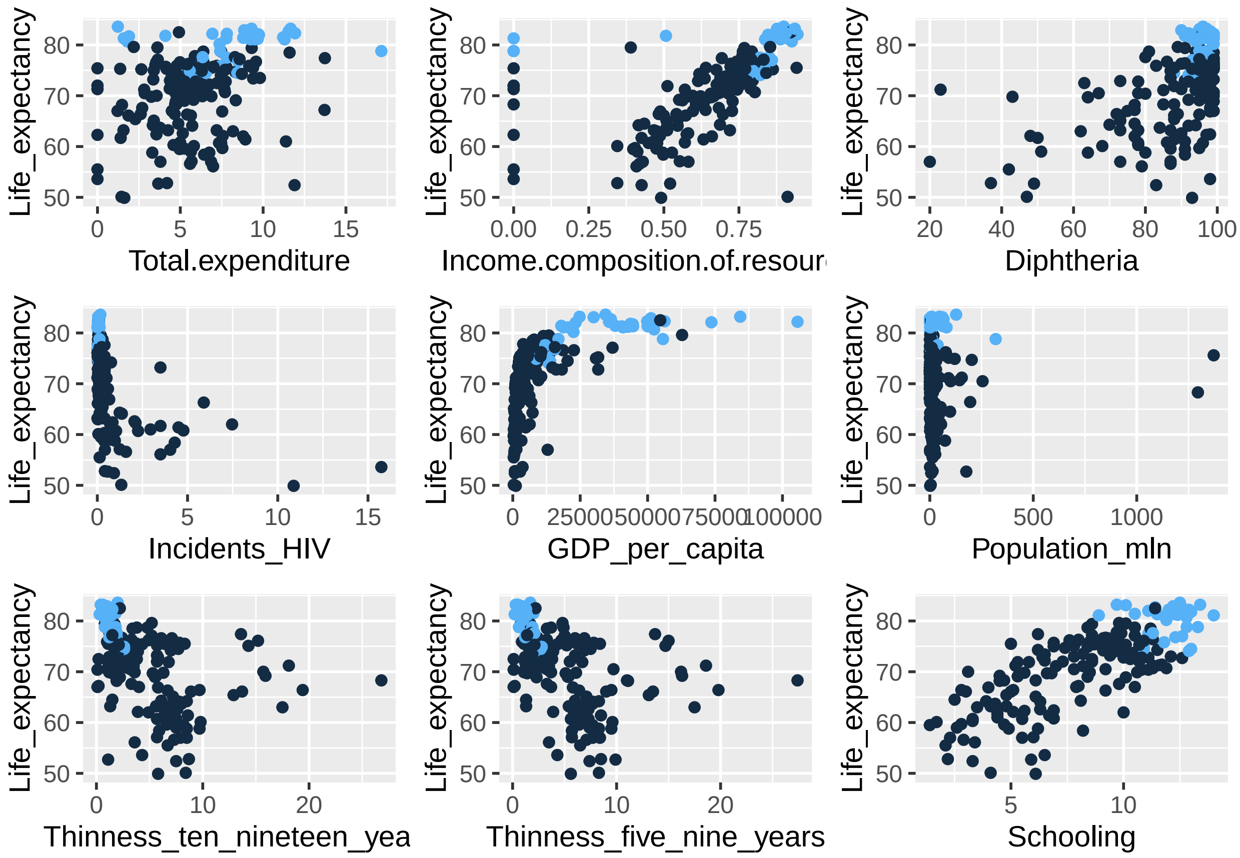
- Developed and developing : The scatterplots contrast life expectancy with various health-related indicators, distinguishing between developed (likely in blue) and developing countries (likely in black). Generally, developed countries show higher life expectancy and lower rates of adult mortality, infant deaths, and under-five deaths. These countries also have higher Hepatitis B and polio immunization rates, higher health expenditure percentages, and lower incidence of measles. Body Mass Index (BMI) distributions appear similar across developed and developing countries. Alcohol consumption does not display a distinct pattern relative to development status. The scatterplots also highlight the differences in life expectancy between developed and developing countries, showing that developed countries often have higher life expectancy associated with higher GDP per capita, education, and immunization rates, and lower HIV incidents and prevalence of thinness among children and adolescents.

```
gg1 <- ggplot(data=led) + geom_point(aes(x=Adult_mortality, y=Life_expectancy, color = Economy_status_De
  theme(legend.position = "none")
gg2 <- ggplot(data=led) + geom_point(aes(x=Infant_deaths, y=Life_expectancy, color =Economy_status_Devel
  theme(legend.position = "none")
gg3 <- ggplot(data=led) + geom_point(aes(x=Alcohol_consumption, y=Life_expectancy,color =Economy_status_
  theme(legend.position = "none")
gg4 <- ggplot(data=led) + geom_point(aes(x=percentage.expenditure, y=Life_expectancy,color =Economy_sta
  theme(legend.position = "none")
gg5 <- ggplot(data=led) + geom_point(aes(x=Hepatitis_B, y=Life_expectancy,color =Economy_status_Develope
  theme(legend.position = "none")
gg6 <- ggplot(data=led) + geom_point(aes(x=Measles, y=Life_expectancy,color =Economy_status_Developed))
  theme(legend.position = "none")
```

```r
gg7 <- ggplot(data=led) + geom_point(aes(x=BMI, y=Life_expectancy,color =Economy_status_Developed)) +
  theme(legend.position = "none")
gg8 <- ggplot(data=led) + geom_point(aes(x=Under_five_deaths, y=Life_expectancy,color =Economy_status_De
  theme(legend.position = "none")
gg9 <- ggplot(data=led) + geom_point(aes(x=Polio, y=Life_expectancy,color =Economy_status_Developed)) +
  theme(legend.position = "none")
gg10 <- ggplot(data=led) + geom_point(aes(x=Total.expenditure, y=Life_expectancy,color =Economy_status_
  theme(legend.position = "none")
gg11 <- ggplot(data=led) + geom_point(aes(x=Income.composition.of.resources, y=Life_expectancy,color =E
  theme(legend.position = "none")
gg12 <- ggplot(data=led) + geom_point(aes(x=Diphtheria, y=Life_expectancy, color =Economy_status_Develo
  theme(legend.position = "none")
gg13 <- ggplot(data=led) + geom_point(aes(x=Incidents_HIV, y=Life_expectancy, color =Economy_status_Dev
  theme(legend.position = "none")
gg14 <- ggplot(data=led) + geom_point(aes(x=GDP_per_capita, y=Life_expectancy, color =Economy_status_De
  theme(legend.position = "none")
gg15 <- ggplot(data=led) + geom_point(aes(x=Population_mln, y=Life_expectancy, color =Economy_status_De
  theme(legend.position = "none")
gg16 <- ggplot(data=led) + geom_point(aes(x=Thinness_ten_nineteen_years, y=Life_expectancy, color =Econ
  theme(legend.position = "none")
gg17 <- ggplot(data=led) + geom_point(aes(x=Thinness_five_nine_years, y=Life_expectancy, color =Economy
  theme(legend.position = "none")
gg18 <- ggplot(data=led) + geom_point(aes(x=Schooling, y=Life_expectancy, color =Economy_status_Develope
  theme(legend.position = "none")

grid.arrange(gg1, gg2, gg3, gg4, gg5, gg6, gg7, gg8, gg9, ncol=3)
```



```r
grid.arrange(gg10, gg11, gg12, gg13, gg14, gg15, gg16, gg17, gg18, ncol=3)
```



```r
names(led)
```

```
##  [1] "Country"                         "Region"
##  [3] "Economy_status_Developed"        "Life_expectancy"
##  [5] "Adult_mortality"                 "Infant_deaths"
##  [7] "Alcohol_consumption"             "percentage.expenditure"
##  [9] "Hepatitis_B"                     "Measles"
## [11] "BMI"                             "Under_five_deaths"
## [13] "Polio"                           "Total.expenditure"
## [15] "Income.composition.of.resources" "Diphtheria"
## [17] "Incidents_HIV"                   "GDP_per_capita"
## [19] "Population_mln"                  "Thinness_ten_nineteen_years"
## [21] "Thinness_five_nine_years"        "Schooling"
```

```r
led %>%
  group_by(Economy_status_Developed) %>%
  select(Country) %>%
  summarise(n())
```

```
## Adding missing grouping variables: ‘Economy_status_Developed‘

## # A tibble: 2 x 2
##   Economy_status_Developed ‘n()‘
##                      <int> <int>
## 1                        0   141
## 2                        1    37
```

2.3 generate hypothesis.

$$H0 := \beta_1 = \beta_2 = ... = \beta_q = 0$$

$$H1 : \text{at least one } \beta_j \neq 0,\ q < j \leq p$$

3. Model building

3.1 Statistical models

3.2 Test Hypothesis

3.3 Predictive model

4. Model diagnostic, selection and comparison

- List all the model assumptions and check if they are violated (e.g. residual plots)
- compare different candidate models
- select the important variables.

Multi Linear Regresssion

```
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
set.seed(123) # For reproducibility
partition <- createDataPartition(led$Life_expectancy, p = 0.8, list = FALSE)
trainingSet <- led[partition, ]
testSet <- led[-partition, ]
```

```
#Multi Linear Regression Model
mlrModel <- lm(Life_expectancy ~ Economy_status_Developed + Adult_mortality + Infant_deaths
               + Alcohol_consumption + percentage.expenditure + Hepatitis_B + Measles + BMI + Under_five
```

```
# Proceed with prediction
testSet$predictions <- predict(mlrModel, newdata = testSet)
summary(mlrModel)
```

```
##
## Call:
## lm(formula = Life_expectancy ~ Economy_status_Developed + Adult_mortality +
##     Infant_deaths + Alcohol_consumption + percentage.expenditure +
##     Hepatitis_B + Measles + BMI + Under_five_deaths + Polio +
##     Total.expenditure + Income.composition.of.resources + Diphtheria +
##     Incidents_HIV + GDP_per_capita + Population_mln + Thinness_ten_nineteen_years +
##     Thinness_five_nine_years + Schooling, data = trainingSet)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0742 -0.9057 -0.1632  0.9466  3.3491
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      8.023e+01  1.422e+00  56.420  < 2e-16 ***
## Economy_status_Developed         8.476e-01  4.841e-01   1.751  0.08250 .
## Adult_mortality                 -5.691e-02  4.007e-03 -14.203  < 2e-16 ***
## Infant_deaths                   -1.000e-01  1.664e-02  -6.012 1.93e-08 ***
## Alcohol_consumption              1.352e-01  4.922e-02   2.747  0.00691 **
## percentage.expenditure          -5.306e-05  7.476e-05  -0.710  0.47922
## Hepatitis_B                      7.165e-03  5.367e-03   1.335  0.18433
## Measles                          3.964e-05  1.905e-05   2.081  0.03948 *
## BMI                             -1.083e-02  7.098e-03  -1.526  0.12945
## Under_five_deaths               -3.956e-03  2.120e-03  -1.866  0.06442 .
## Polio                           -1.631e-02  6.734e-03  -2.423  0.01686 *
## Total.expenditure                6.652e-02  4.364e-02   1.524  0.12998
## Income.composition.of.resources  2.501e+00  7.665e-01   3.263  0.00143 **
## Diphtheria                       9.784e-03  1.106e-02   0.885  0.37794
## Incidents_HIV                    2.245e-01  9.356e-02   2.400  0.01789 *
## GDP_per_capita                   3.293e-05  1.453e-05   2.266  0.02519 *
## Population_mln                   2.176e-03  2.383e-03   0.913  0.36296
## Thinness_ten_nineteen_years     -4.584e-02  9.073e-02  -0.505  0.61431
## Thinness_five_nine_years         2.489e-02  9.277e-02   0.268  0.78889
## Schooling                       -7.292e-03  8.428e-02  -0.087  0.93119
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.363 on 123 degrees of freedom
## Multiple R-squared:  0.9756, Adjusted R-squared:  0.9718
## F-statistic: 258.5 on 19 and 123 DF,  p-value: < 2.2e-16
```

```r
# Calculate evaluation metrics
MAE <- mean(abs(testSet$predictions - testSet$Life_expectancy))
MSE <- mean((testSet$predictions - testSet$Life_expectancy)^2)
RMSE <- sqrt(MSE)

# Print the metrics
print(paste("MAE:", MAE))
```

```
## [1] "MAE: 1.29957748746459"
```

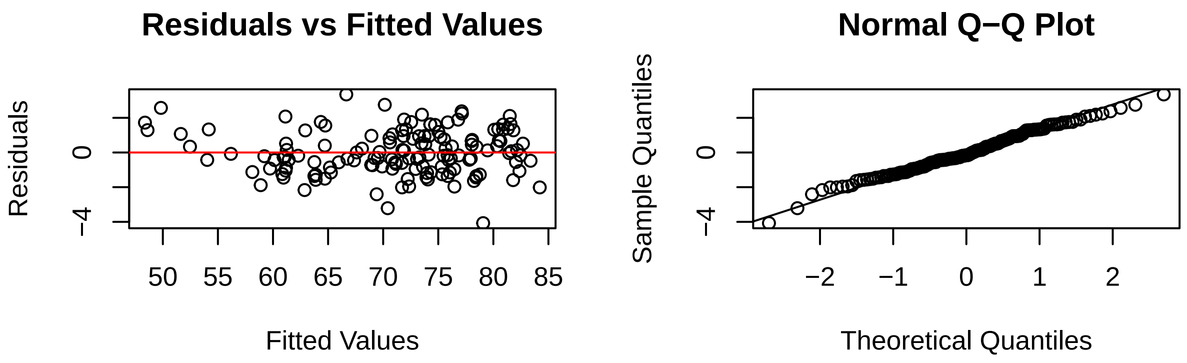
```r
print(paste("MSE:", MSE))
```

```
## [1] "MSE: 2.94264888221154"
```

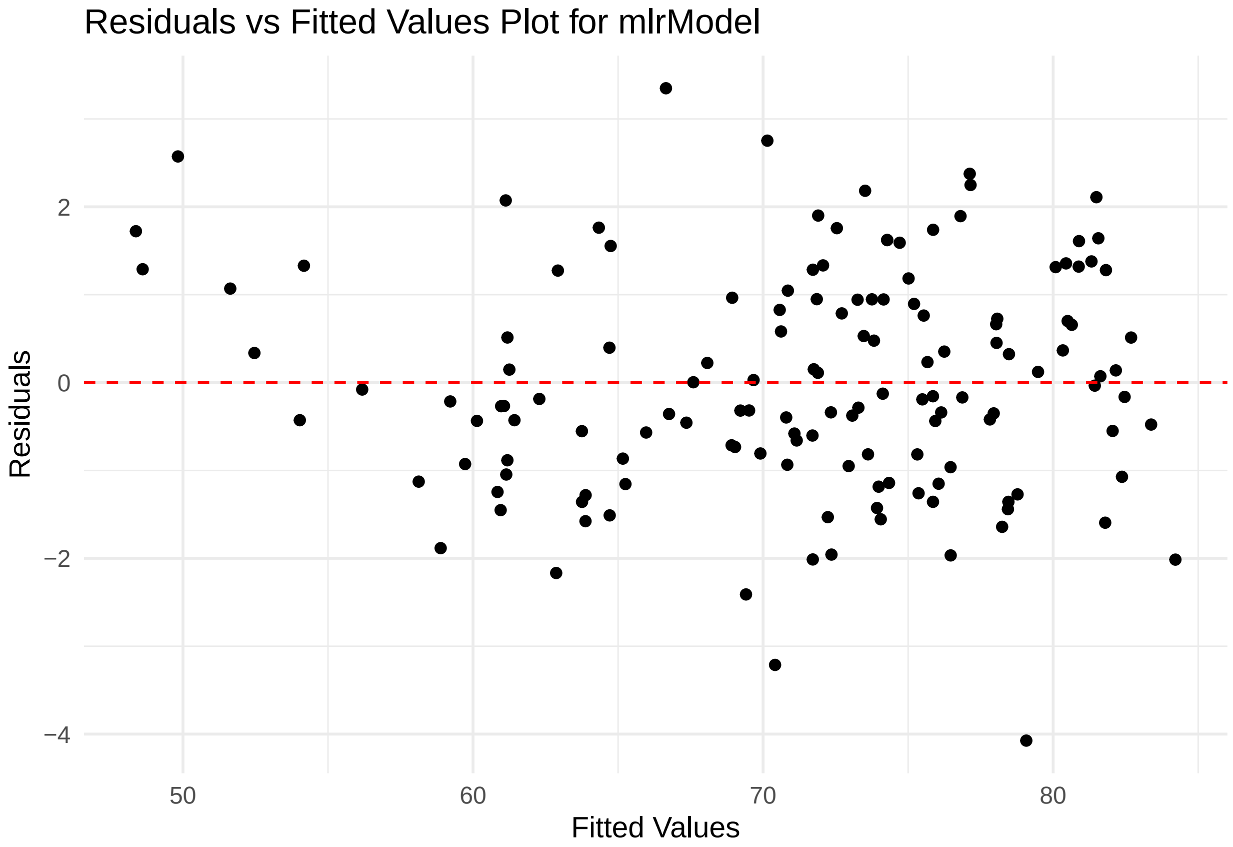
```r
print(paste("RMSE:", RMSE))
```

```
## [1] "RMSE: 1.71541507577949"
```

```r
par(mfrow = c(2, 2)) # Set up the plotting area to display 2 rows of 2 plots each

# Plot 1: Residuals vs Fitted Values
plot(fitted(mlrModel), resid(mlrModel),
     xlab = "Fitted Values",
     ylab = "Residuals",
     main = "Residuals vs Fitted Values")
abline(h = 0, col = "red")

# Plot 2: Normal Q-Q Plot
qqnorm(resid(mlrModel),
       main = "Normal Q-Q Plot")
qqline(resid(mlrModel)) # Add a reference line

# Reset the plotting area back to default (1x1)
par(mfrow = c(1, 1))
```



```r
library(ggplot2)

ggplot(data = data.frame(Fitted = mlrModel$fitted.values, Residuals = resid(mlrModel)), aes(x = Fitted,
  geom_point() +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Fitted Values", y = "Residuals", title = "Residuals vs Fitted Values Plot for mlrModel") +
  theme_minimal()
```



Forward selection

```r
# Minimal model: only intercept (no predictors)
minimalModel <- lm(Life_expectancy ~ 1, data = led)

# Full model: includes all potential predictors
fullModel <- lm(Life_expectancy ~  Economy_status_Developed + Adult_mortality + Infant_deaths
                + Alcohol_consumption + percentage.expenditure + Hepatitis_B + Measles + BMI + Under_fiv
```

```r
# Display the selected model
summary(minimalModel)
```

```
##
## Call:
```

```
## lm(formula = Life_expectancy ~ 1, data = led)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.405  -5.155   1.495   5.370  12.295
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  71.3051     0.5989   119.1   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 7.991 on 177 degrees of freedom
```

```
# Perform forward selection
forwardModel <- step(object = minimalModel, scope = list(lower = minimalModel, upper = fullModel), dire
```

```
## Start:  AIC=740.86
## Life_expectancy ~ 1
##
##                                   Df Sum of Sq     RSS    AIC
## + Adult_mortality                  1   10205.3  1096.4 327.60
## + Infant_deaths                    1    9750.6  1551.0 389.35
## + Schooling                        1    6623.4  4678.2 585.86
## + GDP_per_capita                   1    4466.6  6835.1 653.35
## + Economy_status_Developed         1    3736.2  7565.5 671.42
## + Income.composition.of.resources  1    3397.8  7903.9 679.21
## + Diphtheria                       1    3158.8  8142.9 684.52
## + Incidents_HIV                    1    2463.5  8838.2 699.10
## + Thinness_five_nine_years         1    2444.7  8857.0 699.48
## + Thinness_ten_nineteen_years      1    2212.2  9089.5 704.09
## + Alcohol_consumption              1    2075.3  9226.4 706.75
## + BMI                              1    2062.8  9238.9 706.99
## + percentage.expenditure           1    1796.3  9505.4 712.05
## + Polio                            1    1475.0  9826.7 717.97
## + Total.expenditure                1    1258.2 10043.5 721.86
## + Under_five_deaths                1     499.1 10802.6 734.82
## <none>                                         11301.7 740.86
## + Hepatitis_B                      1      56.9 11244.8 741.97
## + Measles                          1      16.1 11285.6 742.61
## + Population_mln                   1       2.5 11299.2 742.82
##
## Step:  AIC=327.6
## Life_expectancy ~ Adult_mortality
##
##                                   Df Sum of Sq     RSS    AIC
## + Infant_deaths                    1    607.33  489.04 185.90
## + Schooling                        1    465.03  631.34 231.36
## + Alcohol_consumption              1    376.21  720.16 254.79
## + Incidents_HIV                    1    218.18  878.19 290.10
## + Economy_status_Developed         1    212.20  884.17 291.31
## + Thinness_five_nine_years         1    169.72  926.65 299.66
## + GDP_per_capita                   1    159.58  936.79 301.60
## + Thinness_ten_nineteen_years      1    154.91  941.46 302.49
## + Diphtheria                       1    104.25  992.12 311.82
## + Income.composition.of.resources  1     77.09 1019.28 316.62
## + percentage.expenditure           1     63.26 1033.11 319.02
## + Total.expenditure                1     56.94 1039.43 320.11
## + Under_five_deaths                1     26.60 1069.77 325.23
## + BMI                              1     22.69 1073.68 325.88
## + Polio                            1     13.12 1083.25 327.46
## <none>                                         1096.37 327.60
## + Population_mln                   1      6.74 1089.63 328.50
## + Measles                          1      0.45 1095.92 329.53
## + Hepatitis_B                      1      0.02 1096.35 329.60
##
## Step:  AIC=185.9
## Life_expectancy ~ Adult_mortality + Infant_deaths
##
##                                   Df Sum of Sq    RSS    AIC
## + Economy_status_Developed         1   118.137 370.90 138.68
## + GDP_per_capita                   1   106.997 382.04 143.94
## + Alcohol_consumption              1    84.024 405.01 154.34
## + Schooling                        1    57.070 431.97 165.81
## + percentage.expenditure           1    49.618 439.42 168.85
## + Incidents_HIV                    1    26.900 462.13 177.82
## + Hepatitis_B                      1    16.402 472.63 181.82
## + Total.expenditure                1    14.109 474.93 182.69
## + Thinness_five_nine_years         1    12.929 476.11 183.13
## + Income.composition.of.resources  1    12.915 476.12 183.13
## + Thinness_ten_nineteen_years      1    12.551 476.48 183.27
## + Polio                            1     6.309 482.73 185.59
## <none>                                         489.04 185.90
## + Measles                          1     1.262 487.77 187.44
## + Population_mln                   1     0.643 488.39 187.66
## + BMI                              1     0.256 488.78 187.80
## + Under_five_deaths                1     0.030 489.01 187.88
## + Diphtheria                       1     0.001 489.03 187.90
##
## Step:  AIC=138.68
## Life_expectancy ~ Adult_mortality + Infant_deaths + Economy_status_Developed
##
##                                   Df Sum of Sq    RSS    AIC
## + GDP_per_capita                   1   26.5629 344.34 127.45
## + Incidents_HIV                    1   14.7254 356.17 133.47
## + percentage.expenditure           1   12.9725 357.93 134.34
## + Schooling                        1    9.0023 361.90 136.30
## + Alcohol_consumption              1    8.2849 362.61 136.66
## + Polio                            1    7.8330 363.07 136.88
## + Income.composition.of.resources  1    6.6583 364.24 137.45
## <none>                                         370.90 138.68
## + Measles                          1    3.9549 366.94 138.77
## + Hepatitis_B                      1    2.3602 368.54 139.54
## + Total.expenditure                1    1.0777 369.82 140.16
## + BMI                              1    0.6337 370.26 140.37
## + Thinness_ten_nineteen_years      1    0.5014 370.40 140.44
## + Thinness_five_nine_years         1    0.4748 370.42 140.45
## + Under_five_deaths                1    0.1132 370.78 140.62
## + Population_mln                   1    0.0260 370.87 140.66
## + Diphtheria                       1    0.0208 370.88 140.67
## 
## Step:  AIC=127.45
## Life_expectancy ~ Adult_mortality + Infant_deaths + Economy_status_Developed + 
##     GDP_per_capita
## 
##                                   Df Sum of Sq    RSS    AIC
## + Alcohol_consumption              1    7.7494 336.59 125.40
## + Incidents_HIV                    1    7.3741 336.96 125.60
## + Polio                            1    7.0570 337.28 125.76
## + Measles                          1    5.0485 339.29 126.82
## + Income.composition.of.resources  1    3.9503 340.38 127.40
## <none>                                         344.34 127.45
## + Schooling                        1    3.5055 340.83 127.63
## + Total.expenditure                1    1.2803 343.05 128.79
## + Hepatitis_B                      1    0.7289 343.61 129.07
## + Under_five_deaths                1    0.2490 344.09 129.32
## + BMI                              1    0.1403 344.19 129.38
## + Thinness_ten_nineteen_years      1    0.1016 344.23 129.40
## + percentage.expenditure           1    0.0694 344.27 129.41
## + Thinness_five_nine_years         1    0.0378 344.30 129.43
## + Diphtheria                       1    0.0320 344.30 129.43
## + Population_mln                   1    0.0030 344.33 129.45
## 
## Step:  AIC=125.4
## Life_expectancy ~ Adult_mortality + Infant_deaths + Economy_status_Developed + 
##     GDP_per_capita + Alcohol_consumption
## 
##                                   Df Sum of Sq    RSS    AIC
## + Incidents_HIV                    1    8.0474 328.54 123.09
## + Polio                            1    5.8642 330.72 124.27
## + Income.composition.of.resources  1    4.9027 331.68 124.79
## + Measles                          1    4.7900 331.80 124.85
## <none>                                         336.59 125.40
## + Total.expenditure                1    1.6527 334.93 126.52
## + Schooling                        1    1.2944 335.29 126.71
## + Hepatitis_B                      1    0.4061 336.18 127.18
## + Under_five_deaths                1    0.1658 336.42 127.31
## + BMI                              1    0.1275 336.46 127.33
## + Diphtheria                       1    0.0706 336.52 127.36
## + Thinness_five_nine_years         1    0.0241 336.56 127.39
## + percentage.expenditure           1    0.0073 336.58 127.39
## + Population_mln                   1    0.0065 336.58 127.39
## + Thinness_ten_nineteen_years      1    0.0054 336.58 127.39
## 
## Step:  AIC=123.09
## Life_expectancy ~ Adult_mortality + Infant_deaths + Economy_status_Developed + 
##     GDP_per_capita + Alcohol_consumption + Incidents_HIV
## 
##                                   Df Sum of Sq    RSS    AIC
## + Income.composition.of.resources  1    6.0736 322.46 121.77
## + Measles                          1    5.7662 322.77 121.94
## + Polio                            1    5.2686 323.27 122.21
## <none>                                         328.54 123.09
## + Total.expenditure                  1    2.6331 325.91 123.66
## + Schooling                          1    0.9275 327.61 124.59
## + Hepatitis_B                        1    0.3285 328.21 124.91
## + Under_five_deaths                  1    0.2309 328.31 124.97
## + BMI                                1    0.1593 328.38 125.00
## + percentage.expenditure             1    0.0613 328.48 125.06
## + Diphtheria                         1    0.0275 328.51 125.08
## + Population_mln                     1    0.0112 328.53 125.08
## + Thinness_five_nine_years           1    0.0088 328.53 125.09
## + Thinness_ten_nineteen_years        1    0.0034 328.53 125.09
##
## Step:  AIC=121.77
## Life_expectancy ~ Adult_mortality + Infant_deaths + Economy_status_Developed +
##     GDP_per_capita + Alcohol_consumption + Incidents_HIV + Income.composition.of.resources
##
##                               Df Sum of Sq    RSS    AIC
## + Polio                        1    8.5963 313.87 118.96
## + Measles                      1    5.4679 317.00 120.72
## <none>                                     322.46 121.77
## + Total.expenditure            1    1.8638 320.60 122.74
## + BMI                          1    1.3252 321.14 123.04
## + Hepatitis_B                  1    0.8490 321.62 123.30
## + Under_five_deaths            1    0.2311 322.23 123.64
## + Schooling                    1    0.1607 322.30 123.68
## + percentage.expenditure       1    0.0543 322.41 123.74
## + Thinness_five_nine_years     1    0.0359 322.43 123.75
## + Population_mln               1    0.0153 322.45 123.76
## + Thinness_ten_nineteen_years  1    0.0009 322.46 123.77
## + Diphtheria                   1    0.0002 322.46 123.77
##
## Step:  AIC=118.96
## Life_expectancy ~ Adult_mortality + Infant_deaths + Economy_status_Developed +
##     GDP_per_capita + Alcohol_consumption + Incidents_HIV + Income.composition.of.resources +
##     Polio
##
##                               Df Sum of Sq    RSS    AIC
## + Measles                      1    5.4017 308.47 117.87
## <none>                                     313.87 118.96
## + Total.expenditure            1    3.4323 310.44 119.00
## + Diphtheria                   1    1.4659 312.40 120.13
## + BMI                          1    0.9919 312.88 120.40
## + Hepatitis_B                  1    0.8800 312.99 120.46
## + Thinness_five_nine_years     1    0.5047 313.36 120.67
## + Under_five_deaths            1    0.2796 313.59 120.80
## + percentage.expenditure       1    0.2319 313.64 120.83
## + Thinness_ten_nineteen_years  1    0.2028 313.67 120.84
## + Schooling                    1    0.0261 313.84 120.94
## + Population_mln               1    0.0031 313.87 120.96
##
## Step:  AIC=117.87
## Life_expectancy ~ Adult_mortality + Infant_deaths + Economy_status_Developed +
##     GDP_per_capita + Alcohol_consumption + Incidents_HIV + Income.composition.of.resources +
##     Polio + Measles
```

```
## 
##                                 Df Sum of Sq    RSS    AIC
## + Total.expenditure              1    3.7670 304.70 117.68
## <none>                                       308.47 117.87
## + Under_five_deaths              1    2.0778 306.39 118.67
## + Diphtheria                     1    1.9773 306.49 118.72
## + Population_mln                 1    1.9613 306.50 118.73
## + Hepatitis_B                    1    0.8765 307.59 119.36
## + BMI                            1    0.5013 307.96 119.58
## + percentage.expenditure         1    0.2185 308.25 119.74
## + Schooling                      1    0.0321 308.43 119.85
## + Thinness_ten_nineteen_years    1    0.0271 308.44 119.85
## + Thinness_five_nine_years       1    0.0225 308.44 119.86
## 
## Step:  AIC=117.68
## Life_expectancy ~ Adult_mortality + Infant_deaths + Economy_status_Developed + 
##     GDP_per_capita + Alcohol_consumption + Incidents_HIV + Income.composition.of.resources + 
##     Polio + Measles + Total.expenditure
## 
##                                 Df Sum of Sq    RSS    AIC
## <none>                                       304.70 117.68
## + Population_mln                 1   2.45549 302.24 118.24
## + Diphtheria                     1   2.18415 302.51 118.40
## + Under_five_deaths              1   1.98817 302.71 118.52
## + Hepatitis_B                    1   0.89108 303.81 119.16
## + BMI                            1   0.74077 303.96 119.25
## + percentage.expenditure         1   0.23296 304.47 119.55
## + Schooling                      1   0.11826 304.58 119.61
## + Thinness_five_nine_years       1   0.00293 304.70 119.68
## + Thinness_ten_nineteen_years    1   0.00071 304.70 119.68
```

```
# Display the selected model
summary(forwardModel)
```

```
## 
## Call:
## lm(formula = Life_expectancy ~ Adult_mortality + Infant_deaths + 
##     Economy_status_Developed + GDP_per_capita + Alcohol_consumption + 
##     Incidents_HIV + Income.composition.of.resources + Polio + 
##     Measles + Total.expenditure, data = led)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.6255 -0.7875 -0.0377  0.9512  3.4817 
## 
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                      8.192e+01  6.675e-01 122.734  < 2e-16 ***
## Adult_mortality                 -5.318e-02  3.173e-03 -16.761  < 2e-16 ***
## Infant_deaths                   -1.291e-01  1.144e-02 -11.283  < 2e-16 ***
## Economy_status_Developed         1.015e+00  4.302e-01   2.359  0.01949 *  
## GDP_per_capita                   2.523e-05  8.831e-06   2.857  0.00482 ** 
## Alcohol_consumption              8.797e-02  4.093e-02   2.149  0.03307 *  
## Incidents_HIV                    2.053e-01  8.400e-02   2.444  0.01557 *  
## Income.composition.of.resources  1.187e+00  5.610e-01   2.116  0.03583 *
## Polio                           -1.150e-02  4.869e-03  -2.362  0.01931 *
## Measles                          2.126e-05  1.199e-05   1.773  0.07803 .
## Total.expenditure                5.559e-02  3.869e-02   1.437  0.15262
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.351 on 167 degrees of freedom
## Multiple R-squared:  0.973,  Adjusted R-squared:  0.9714
## F-statistic: 602.7 on 10 and 167 DF,  p-value: < 2.2e-16
```

Ridge Regression

```r
library(glmnet)
```

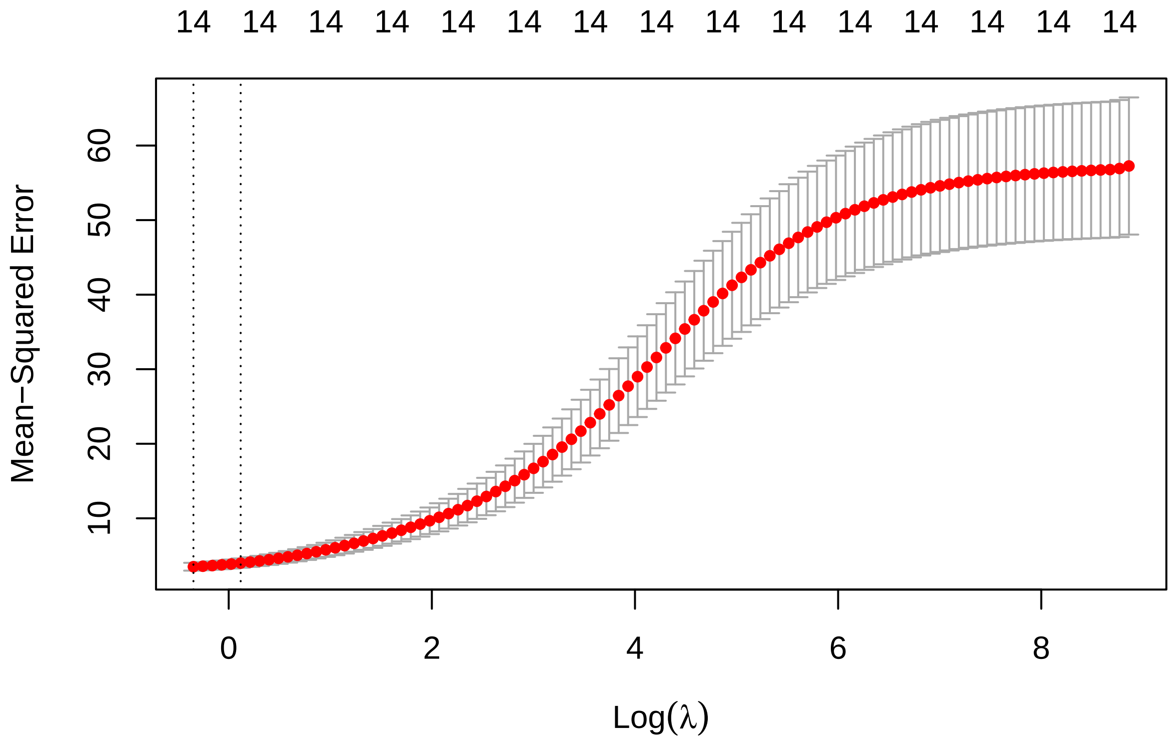
```
## Loading required package: Matrix
```

```
##
## Attaching package: ’Matrix’
```

```
## The following objects are masked from ’package:tidyr’:
##
##     expand, pack, unpack
```

```
## Loaded glmnet 4.1-8
```

```r
subled <- select(led, c(-Country, -Region, -Incidents_HIV, -Under_five_deaths, -Measles, -Infant_deaths

x <- model.matrix(Life_expectancy ~ ., subled)[, -1]
y <- led$Life_expectancy

set.seed(1)
train <- sample(1:nrow(x), nrow(x) / 2)
test <- (-train)
y.test <- y[test]

cv.out <- cv.glmnet(x[train, ], y[train], alpha = 0)
plot(cv.out)
```



```
bestlam <- cv.out$lambda.min
ridge.mod  <- glmnet(x[train,], y[train], alpha=0, lambda=bestlam, thresh = 1e-12)
ridge.pred <- predict(ridge.mod, newx = x[test, ])
mean((ridge.pred - y.test)^2)
```

```
## [1] 4.052225
```

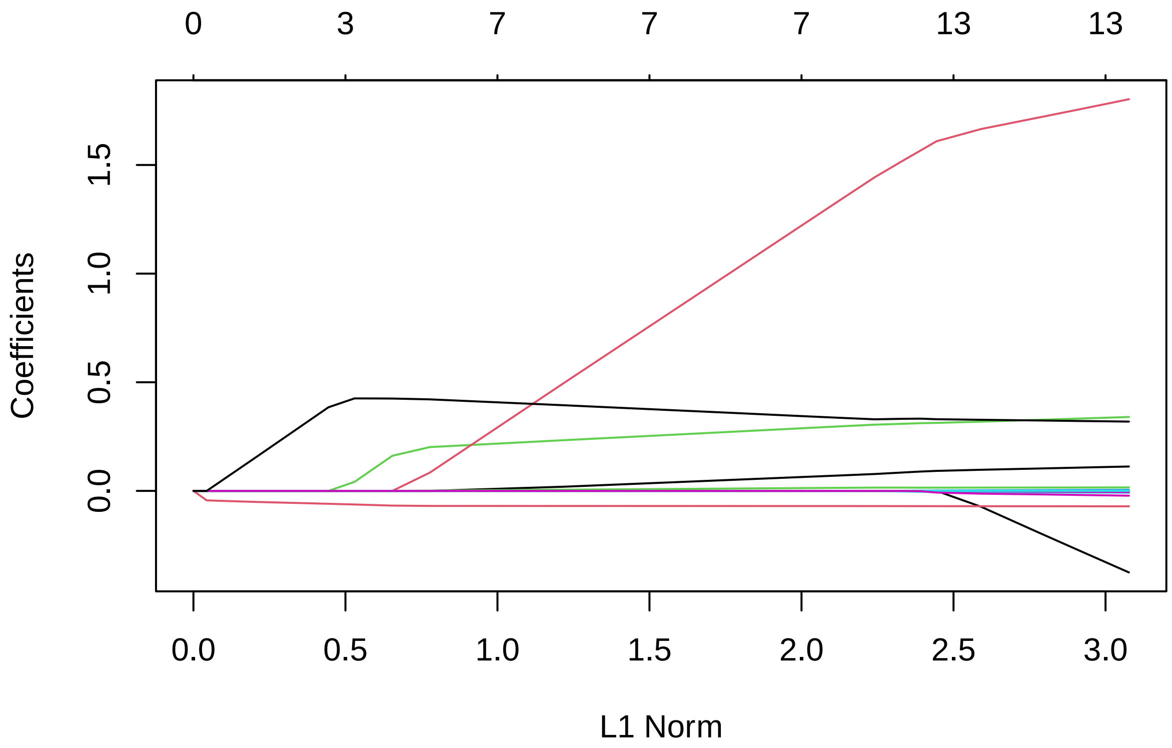
```
out <- glmnet(x, y, alpha = 0)
predict(out, type="coefficients", s = bestlam)
```

```
## 15 x 1 sparse Matrix of class "dgCMatrix"
##                                         s1
## (Intercept)                   7.054910e+01
## Economy_status_Developed      6.427917e-01
## Adult_mortality              -5.613519e-02
## Alcohol_consumption           1.001535e-01
## percentage.expenditure       -2.504167e-05
## Hepatitis_B                   1.966859e-03
## BMI                          -4.616956e-03
## Polio                        -6.970471e-03
## Total.expenditure             8.874715e-02
## Income.composition.of.resources  2.164560e+00
## Diphtheria                    5.088585e-02
## GDP_per_capita                3.003571e-05
## Population_mln                4.215404e-04
## Thinness_five_nine_years        -6.339978e-02
## Schooling                        4.265114e-01
```

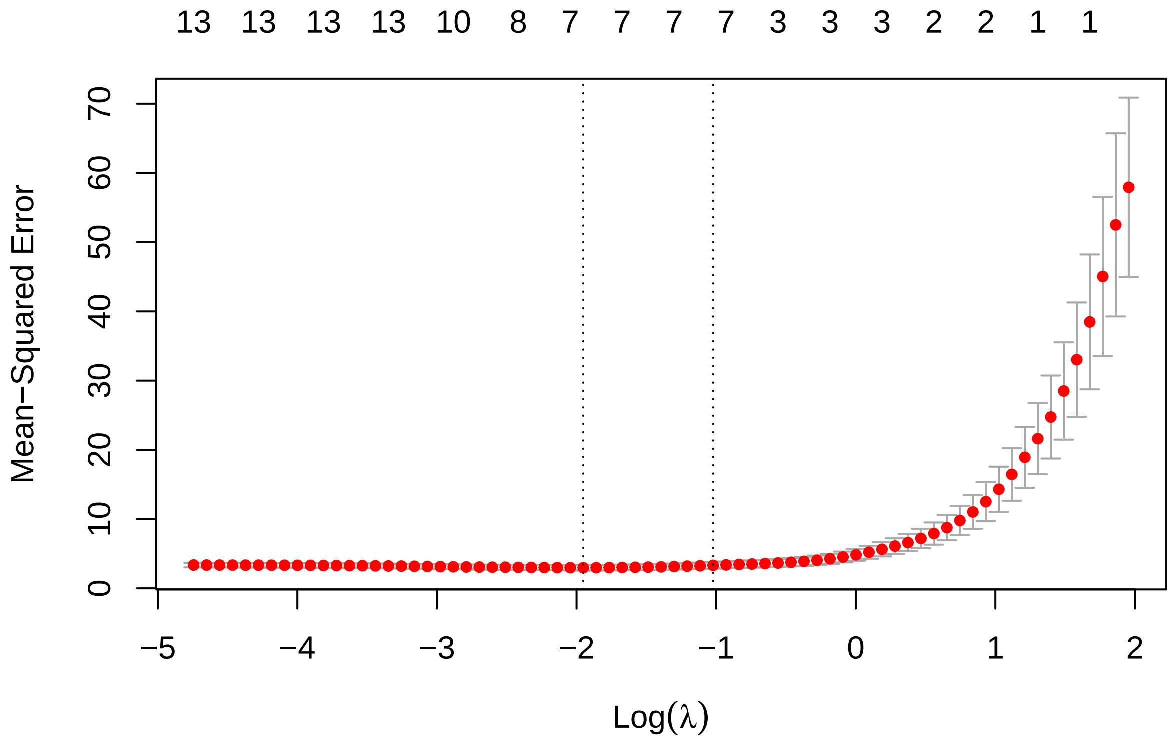
LASSO

```
grid <- 10^seq(10, -2, length = 100)
lasso.mod <- glmnet(x[train, ], y[train], alpha = 1,lambda = grid)
plot(lasso.mod)
```

```
## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values
```



```
set.seed(1)
cv.out <- cv.glmnet(x[train, ], y[train], alpha = 1)
plot(cv.out)
```



```
bestlam <- cv.out$lambda.min
lasso.pred <- predict(lasso.mod, s = bestlam,newx = x[test, ])
mean((lasso.pred - y.test)^2)
```

```
## [1] 3.79647
```

```
out <- glmnet(x, y, alpha=1, lambda = grid)
lasso.coef <- predict(out, type="coefficients", s=bestlam)
lasso.coef
```

```
## 15 x 1 sparse Matrix of class "dgCMatrix"
##                                            s1
## (Intercept)                      7.471679e+01
## Economy_status_Developed         9.229752e-02
## Adult_mortality                 -6.601461e-02
## Alcohol_consumption              1.422658e-01
## percentage.expenditure           .
## Hepatitis_B                      .
## BMI                              .
## Polio                            .
## Total.expenditure                4.939553e-02
## Income.composition.of.resources  7.596336e-01
## Diphtheria                       2.987331e-02
## GDP_per_capita                   1.542573e-05
## Population_mln                   .
## Thinness_five_nine_years        -2.208760e-02
## Schooling                        3.955407e-01
```

```
model.lasso <- glm(Life_expectancy~Income.composition.of.resources+Schooling+Alcohol_consumption+Econom

summary(model.lasso)
```

```
##
## Call:
## glm(formula = Life_expectancy ~ Income.composition.of.resources +
##     Schooling + Alcohol_consumption + Economy_status_Developed +
##     Adult_mortality + Total.expenditure + Diphtheria + Thinness_five_nine_years +
##     GDP_per_capita, data = led)
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      7.408e+01  1.311e+00  56.522  < 2e-16 ***
## Income.composition.of.resources  1.119e+00  7.213e-01   1.551 0.122833
## Schooling                        3.636e-01  7.851e-02   4.632 7.22e-06 ***
## Alcohol_consumption              1.678e-01  5.379e-02   3.120 0.002127 **
## Economy_status_Developed         4.921e-02  5.408e-01   0.091 0.927610
## Adult_mortality                 -6.642e-02  2.189e-03 -30.335  < 2e-16 ***
## Total.expenditure                7.422e-02  4.889e-02   1.518 0.130864
## Diphtheria                       3.575e-02  1.031e-02   3.467 0.000669 ***
## Thinness_five_nine_years        -3.839e-02  3.828e-02  -1.003 0.317453
## GDP_per_capita                   1.812e-05  1.113e-05   1.628 0.105479
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for gaussian family taken to be 2.997509)
##
##     Null deviance: 11301.69  on 177  degrees of freedom
## Residual deviance:   503.58  on 168  degrees of freedom
## AIC: 712.26
##
## Number of Fisher Scoring iterations: 2
```