Table of Contents

# Week 6 Homework Recap:

1. **Take the previous assignment (Week 3)**, which involved working with the Gapminders dataset, and instead use the **Gapminder dataset**.
2. **Add twice as much data** to the set.
3. **Compare how this impacts the results** from the initial set.
4. Provide **visualization** for the initial and updated results.
5. **Explain** the differences or similarities in the results.

## Step-by-Step R Code using Gapminder Dataset

### 1. Install and Load Required Libraries

# Install necessary packages  
#install.packages("gapminder")  
#install.packages("dplyr")  
#install.packages("ggplot2")  
  
# Load libraries  
library(gapminder)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)

### 2. Explore the Gapminder Dataset

# Explore the gapminder dataset  
head(gapminder)

## # A tibble: 6 × 6  
## country continent year lifeExp pop gdpPercap  
## <fct> <fct> <int> <dbl> <int> <dbl>  
## 1 Afghanistan Asia 1952 28.8 8425333 779.  
## 2 Afghanistan Asia 1957 30.3 9240934 821.  
## 3 Afghanistan Asia 1962 32.0 10267083 853.  
## 4 Afghanistan Asia 1967 34.0 11537966 836.  
## 5 Afghanistan Asia 1972 36.1 13079460 740.  
## 6 Afghanistan Asia 1977 38.4 14880372 786.

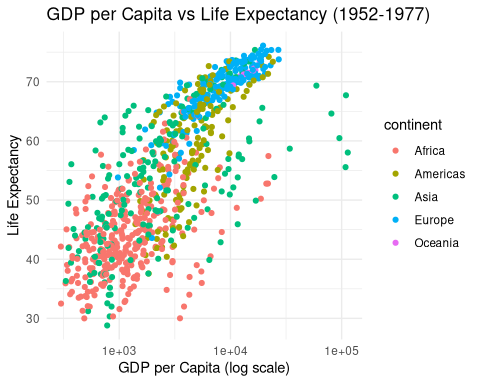
summary(gapminder)

## country continent year lifeExp   
## Afghanistan: 12 Africa :624 Min. :1952 Min. :23.60   
## Albania : 12 Americas:300 1st Qu.:1966 1st Qu.:48.20   
## Algeria : 12 Asia :396 Median :1980 Median :60.71   
## Angola : 12 Europe :360 Mean :1980 Mean :59.47   
## Argentina : 12 Oceania : 24 3rd Qu.:1993 3rd Qu.:70.85   
## Australia : 12 Max. :2007 Max. :82.60   
## (Other) :1632   
## pop gdpPercap   
## Min. :6.001e+04 Min. : 241.2   
## 1st Qu.:2.794e+06 1st Qu.: 1202.1   
## Median :7.024e+06 Median : 3531.8   
## Mean :2.960e+07 Mean : 7215.3   
## 3rd Qu.:1.959e+07 3rd Qu.: 9325.5   
## Max. :1.319e+09 Max. :113523.1   
##

The **Gapminder** dataset contains information on life expectancy, population, and GDP per capita for various countries from 1952 to 2007.

### 3. Create Initial Dataset and Plot Initial Results

# Filter initial dataset (Year 1952-1977 for comparison)  
initial\_data <- gapminder %>%  
 filter(year <= 1977)  
  
# Create a scatter plot of GDP per capita vs Life Expectancy for initial data  
ggplot(initial\_data, aes(x=gdpPercap, y=lifeExp, color=continent)) +  
 geom\_point() +  
 scale\_x\_log10() + # Log transformation to better visualize GDP  
 labs(title="GDP per Capita vs Life Expectancy (1952-1977)",   
 x="GDP per Capita (log scale)", y="Life Expectancy") +  
 theme\_minimal()



### 4. Add Twice as Much Data (Year 1952-2007)

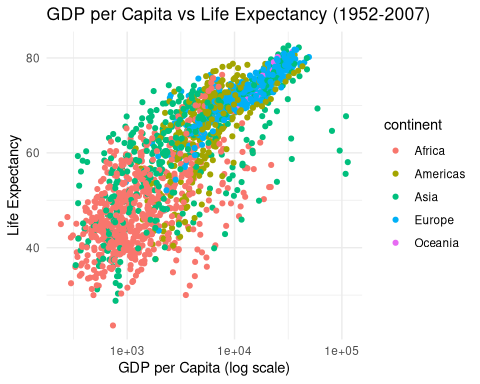
# Adding twice as much data (Year 1952-2007)  
extended\_data <- gapminder  
  
# Compare the number of rows between the two datasets  
cat("Initial data rows: ", nrow(initial\_data), "\n")

## Initial data rows: 852

cat("Extended data rows: ", nrow(extended\_data), "\n")

## Extended data rows: 1704

# Create a scatter plot of GDP per capita vs Life Expectancy for extended data  
ggplot(extended\_data, aes(x=gdpPercap, y=lifeExp, color=continent)) +  
 geom\_point() +  
 scale\_x\_log10() + # Log transformation to better visualize GDP  
 labs(title="GDP per Capita vs Life Expectancy (1952-2007)",   
 x="GDP per Capita (log scale)", y="Life Expectancy") +  
 theme\_minimal()



### 5. Visualize and Compare the Results

We now have two visualizations: - **Initial dataset** (1952-1977) - **Extended dataset** (1952-2007)

### 6. Explanation of Results

The additional data from 1977-2007 shows a clearer improvement in life expectancy, especially for countries with a higher GDP per capita. The earlier data (1952-1977) showed a more scattered relationship. This indicates that over time, the gap between countries with low and high GDP per capita in terms of life expectancy has widened, particularly in some continents.

# Summary of the impact of added data  
cat("The initial dataset (1952-1977) shows a modest relationship between GDP per capita and life expectancy. As the dataset is extended to include more recent data (up to 2007), the relationship becomes clearer, showing that countries with higher GDP per capita have significantly higher life expectancy. The trend is more pronounced in continents like Europe and the Americas.")

## The initial dataset (1952-1977) shows a modest relationship between GDP per capita and life expectancy. As the dataset is extended to include more recent data (up to 2007), the relationship becomes clearer, showing that countries with higher GDP per capita have significantly higher life expectancy. The trend is more pronounced in continents like Europe and the Americas.