Data Exploration and Visualization with Gapminder

Mark

2025-03-21

# Introduction

This project explores the gapminder dataset using **R** and **tidyverse**. We perform **data cleaning, descriptive analysis, and visualizations**, along with statistical tests such as **t-tests** and **ANOVA**.

# Installing and Loading Packages

install.packages("rmarkdown")  
install.packages("shiny")

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.4.2

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.4.2

## Warning: package 'tibble' was built under R version 4.4.2

## Warning: package 'tidyr' was built under R version 4.4.2

## Warning: package 'readr' was built under R version 4.4.2

## Warning: package 'purrr' was built under R version 4.4.2

## Warning: package 'dplyr' was built under R version 4.4.2

## Warning: package 'stringr' was built under R version 4.4.2

## Warning: package 'forcats' was built under R version 4.4.2

## Warning: package 'lubridate' was built under R version 4.4.2

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ lubridate 1.9.4 ✔ tibble 3.2.1  
## ✔ purrr 1.0.4 ✔ tidyr 1.3.1  
## ── 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

library(gapminder)

## Warning: package 'gapminder' was built under R version 4.4.2

# Data Exploration

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.

tail(gapminder)

## # A tibble: 6 × 6  
## country continent year lifeExp pop gdpPercap  
## <fct> <fct> <int> <dbl> <int> <dbl>  
## 1 Zimbabwe Africa 1982 60.4 7636524 789.  
## 2 Zimbabwe Africa 1987 62.4 9216418 706.  
## 3 Zimbabwe Africa 1992 60.4 10704340 693.  
## 4 Zimbabwe Africa 1997 46.8 11404948 792.  
## 5 Zimbabwe Africa 2002 40.0 11926563 672.  
## 6 Zimbabwe Africa 2007 43.5 12311143 470.

glimpse(gapminder)

## Rows: 1,704  
## Columns: 6  
## $ country <fct> "Afghanistan", "Afghanistan", "Afghanistan", "Afghanistan", …  
## $ continent <fct> Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, …  
## $ year <int> 1952, 1957, 1962, 1967, 1972, 1977, 1982, 1987, 1992, 1997, …  
## $ lifeExp <dbl> 28.801, 30.332, 31.997, 34.020, 36.088, 38.438, 39.854, 40.8…  
## $ pop <int> 8425333, 9240934, 10267083, 11537966, 13079460, 14880372, 12…  
## $ gdpPercap <dbl> 779.4453, 820.8530, 853.1007, 836.1971, 739.9811, 786.1134, …

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   
##

unique(gapminder$country)

## [1] Afghanistan Albania Algeria   
## [4] Angola Argentina Australia   
## [7] Austria Bahrain Bangladesh   
## [10] Belgium Benin Bolivia   
## [13] Bosnia and Herzegovina Botswana Brazil   
## [16] Bulgaria Burkina Faso Burundi   
## [19] Cambodia Cameroon Canada   
## [22] Central African Republic Chad Chile   
## [25] China Colombia Comoros   
## [28] Congo, Dem. Rep. Congo, Rep. Costa Rica   
## [31] Cote d'Ivoire Croatia Cuba   
## [34] Czech Republic Denmark Djibouti   
## [37] Dominican Republic Ecuador Egypt   
## [40] El Salvador Equatorial Guinea Eritrea   
## [43] Ethiopia Finland France   
## [46] Gabon Gambia Germany   
## [49] Ghana Greece Guatemala   
## [52] Guinea Guinea-Bissau Haiti   
## [55] Honduras Hong Kong, China Hungary   
## [58] Iceland India Indonesia   
## [61] Iran Iraq Ireland   
## [64] Israel Italy Jamaica   
## [67] Japan Jordan Kenya   
## [70] Korea, Dem. Rep. Korea, Rep. Kuwait   
## [73] Lebanon Lesotho Liberia   
## [76] Libya Madagascar Malawi   
## [79] Malaysia Mali Mauritania   
## [82] Mauritius Mexico Mongolia   
## [85] Montenegro Morocco Mozambique   
## [88] Myanmar Namibia Nepal   
## [91] Netherlands New Zealand Nicaragua   
## [94] Niger Nigeria Norway   
## [97] Oman Pakistan Panama   
## [100] Paraguay Peru Philippines   
## [103] Poland Portugal Puerto Rico   
## [106] Reunion Romania Rwanda   
## [109] Sao Tome and Principe Saudi Arabia Senegal   
## [112] Serbia Sierra Leone Singapore   
## [115] Slovak Republic Slovenia Somalia   
## [118] South Africa Spain Sri Lanka   
## [121] Sudan Swaziland Sweden   
## [124] Switzerland Syria Taiwan   
## [127] Tanzania Thailand Togo   
## [130] Trinidad and Tobago Tunisia Turkey   
## [133] Uganda United Kingdom United States   
## [136] Uruguay Venezuela Vietnam   
## [139] West Bank and Gaza Yemen, Rep. Zambia   
## [142] Zimbabwe   
## 142 Levels: Afghanistan Albania Algeria Angola Argentina Australia ... Zimbabwe

class(gapminder$pop)

## [1] "integer"

# Data Cleaning

noMissingData <- drop\_na(gapminder)  
missingData <- !complete.cases(gapminder)  
gapminder[missingData,]

## # A tibble: 0 × 6  
## # ℹ 6 variables: country <fct>, continent <fct>, year <int>, lifeExp <dbl>,  
## # pop <int>, gdpPercap <dbl>

cleaned\_data <- gapminder %>%   
 select(country, pop, year) %>%   
 rename(population = pop) %>%   
 filter(year > 2000 & population > 10000000) %>%   
 arrange(year)  
head(cleaned\_data)

## # A tibble: 6 × 3  
## country population year  
## <fct> <int> <int>  
## 1 Afghanistan 25268405 2002  
## 2 Algeria 31287142 2002  
## 3 Angola 10866106 2002  
## 4 Argentina 38331121 2002  
## 5 Australia 19546792 2002  
## 6 Bangladesh 135656790 2002

sampleData <- select(gapminder, country, pop, year)  
yearly\_data <- sampleData %>%   
 pivot\_wider(names\_from = year, values\_from = pop)  
View(yearly\_data)

# Descriptive Statistics

population <- gapminder$pop  
summary(population)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 6.001e+04 2.794e+06 7.024e+06 2.960e+07 1.959e+07 1.319e+09

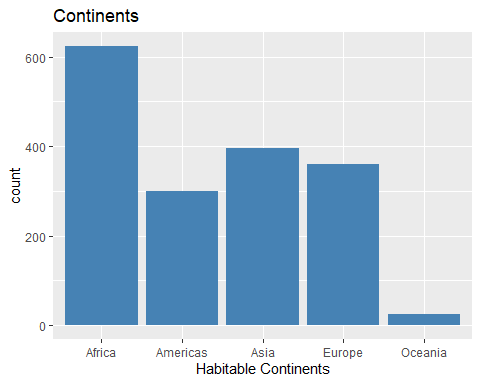
population <- population / 1000000  
summary(population)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.060 2.794 7.024 29.601 19.585 1318.683

# Data Visualization

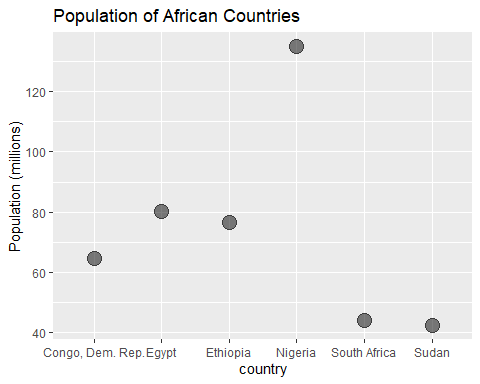
## Bar Chart of Continents

ggplot(data=gapminder, aes(x=continent)) +   
 geom\_bar(fill = "steelblue") +   
 labs(title = "Continents", x="Habitable Continents")



## Scatter Plot of Population (African Countries)

gapminder %>%   
 filter(year > 2006 & continent == "Africa" & pop > 40000000) %>%   
 ggplot(aes(x = country, y = pop/1000000)) +   
 geom\_point(size=5, alpha = 0.5) +  
 labs(title = "Population of African Countries", y="Population (millions)")



# Statistical Analysis

## T-test (Life Expectancy in Americas vs Oceania)

gapminder %>%   
 filter(continent %in% c("Americas", "Oceania")) %>%   
 t.test(lifeExp ~ continent, data = .)

##   
## Welch Two Sample t-test  
##   
## data: lifeExp by continent  
## t = -10.24, df = 49.815, p-value = 7.523e-14  
## alternative hypothesis: true difference in means between group Americas and group Oceania is not equal to 0  
## 95 percent confidence interval:  
## -11.563985 -7.770958  
## sample estimates:  
## mean in group Americas mean in group Oceania   
## 64.65874 74.32621

## ANOVA (Life Expectancy Across Continents)

gapminder %>%   
 filter(year == 2007, continent %in% c("Asia", "Africa", "Americas")) %>%   
 aov(lifeExp ~ continent, data = .) %>%  
 summary()

## Df Sum Sq Mean Sq F value Pr(>F)   
## continent 2 8195 4098 60.62 <2e-16 \*\*\*  
## Residuals 107 7233 68   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Conclusion

This project explored the gapminder dataset, cleaned the data, visualized key insights, and performed statistical tests. The findings suggest patterns in population distribution and life expectancy differences among continents.