

GroupProject

Grace Rovinsky, Max Lick, Jeremy Ijoma

2025-12-17

Introduction

Understanding how demographic, economic, and health variables interact with each other is essential for evaluating the development of countries and informing future policy. This project uses the countryData dataset from the dcData package to conduct an exploratory data analysis of countries' attributes such as population, education, public health, and macroeconomic conditions. We used data visualizations to identify patterns and relationships across multiple variables within our data. By displaying multivariate data visualizations and descriptive statistics using the tidyverse and ggplot2 packages, we highlighted how complex attributes such as population growth, economic stability, and health are shaped by many connected factors.

Summary Table

This table includes a summary of the descriptive statistics we chose to look at for this project.

```
-- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
v dplyr      1.1.4      v readr      2.1.5
v forcats    1.0.1      v stringr    1.5.2
v ggplot2     3.5.2      v tibble     3.3.0
v lubridate  1.9.4      v tidyr      1.3.1
v purrr      1.1.0

-- Conflicts ----- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()     masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
```

Attaching package: 'kableExtra'

The following object is masked from 'package:dplyr':

```
group_rows
```

```
Attaching package: 'rvest'
```

```
The following object is masked from 'package:readr':
```

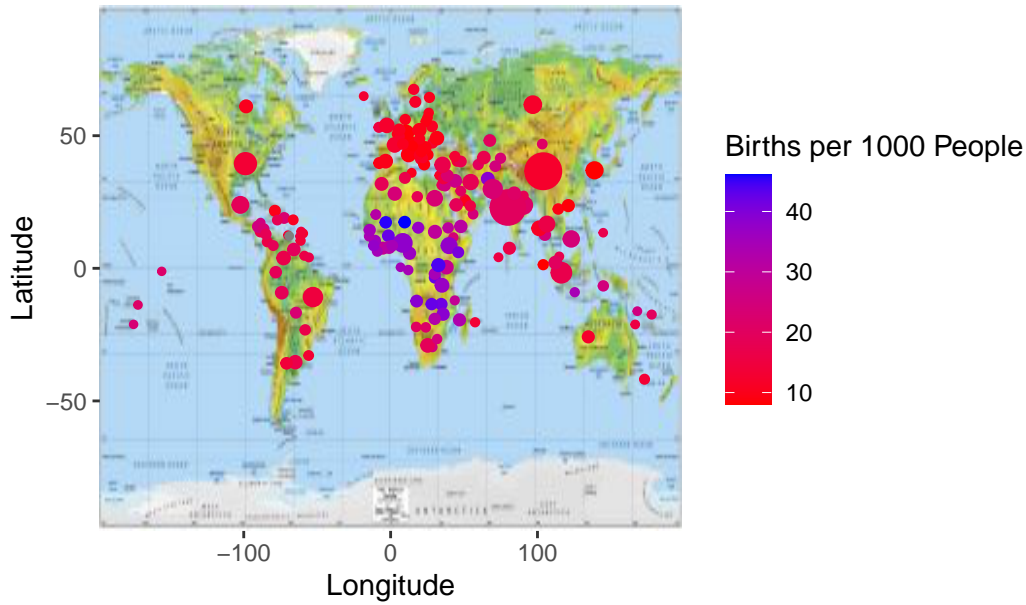
```
guess_encoding
```

	vars	n	mean	sd	median	trimmed	mad
area	1	252	541421.52	1712015.64	56689.50	186116.49	83922.57
pop	2	240	32025370.01	127199862.70	5159867.50	9962552.71	7568898.36
growth	3	233	1.10	1.42	1.02	1.07	1.22
birth	4	224	19.66	9.67	16.89	18.58	9.38
death	5	225	7.91	2.98	7.63	7.71	2.56
health	6	190	7.02	3.08	6.55	6.75	2.89
obesity	7	191	19.34	13.50	20.10	18.09	12.90
educ	8	173	4.85	2.14	4.70	4.67	1.93
unemployment	9	203	14.48	15.48	9.40	11.47	7.86
life	10	223	71.80	8.81	74.33	72.79	7.55
	min		max	range	skew	kurtosis	se
area	0.00	1.709824e+07	1.709824e+07	6.22	45.11	107846.85	
pop	48.00	1.355693e+09	1.355693e+09	8.56	80.37	8210715.83	
growth	-9.73	9.370000e+00	1.910000e+01	-0.96	17.69	0.09	
birth	6.72	4.612000e+01	3.940000e+01	0.86	-0.27	0.65	
death	1.53	1.749000e+01	1.596000e+01	0.57	0.01	0.20	
health	1.90	1.950000e+01	1.760000e+01	1.24	2.60	0.22	
obesity	1.10	7.460000e+01	7.350000e+01	1.12	2.33	0.98	
educ	0.60	1.300000e+01	1.240000e+01	1.08	2.17	0.16	
unemployment	0.00	9.500000e+01	9.500000e+01	2.66	8.66	1.09	
life	49.44	8.957000e+01	4.013000e+01	-0.89	-0.07	0.59	

Population vs. Birth Rate

```
Warning: Removed 25 rows containing missing values or values outside the scale range
(`geom_point()`).
```

Map of Population and Birth Rate



Country Data by Continent

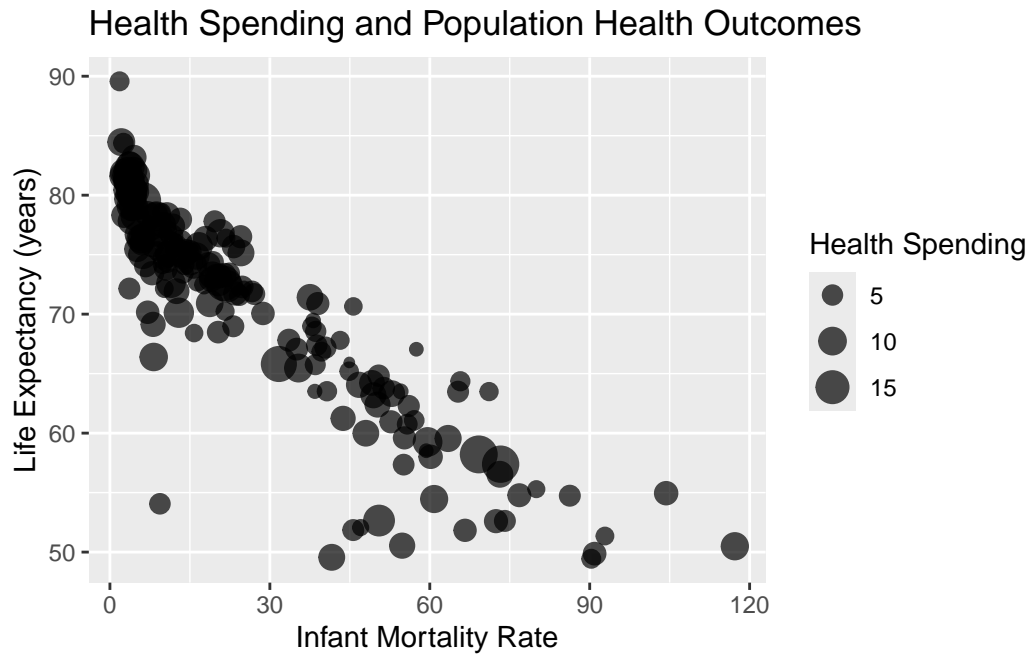
Table 1: Country Data Summary Statistics

Continent	Area_mean	Pop_mean	Growth_mean	Birth_mean	Death_mean
Africa	565264.2	20769380	1.8472000	28.98960	9.66400
Asia	1006834.8	77988744	1.0250000	18.91806	6.19600
Europe	166297.0	16629907	0.1422857	10.44143	10.50800
North America	1118219.4	26788635	0.7615000	16.27350	6.56800
Oceania	893536.3	3154768	0.5544444	18.72778	6.28000
South America	1597057.5	36851369	0.9436364	17.10727	6.31600

Health Spending, Life Expectancy, Infant Mortality

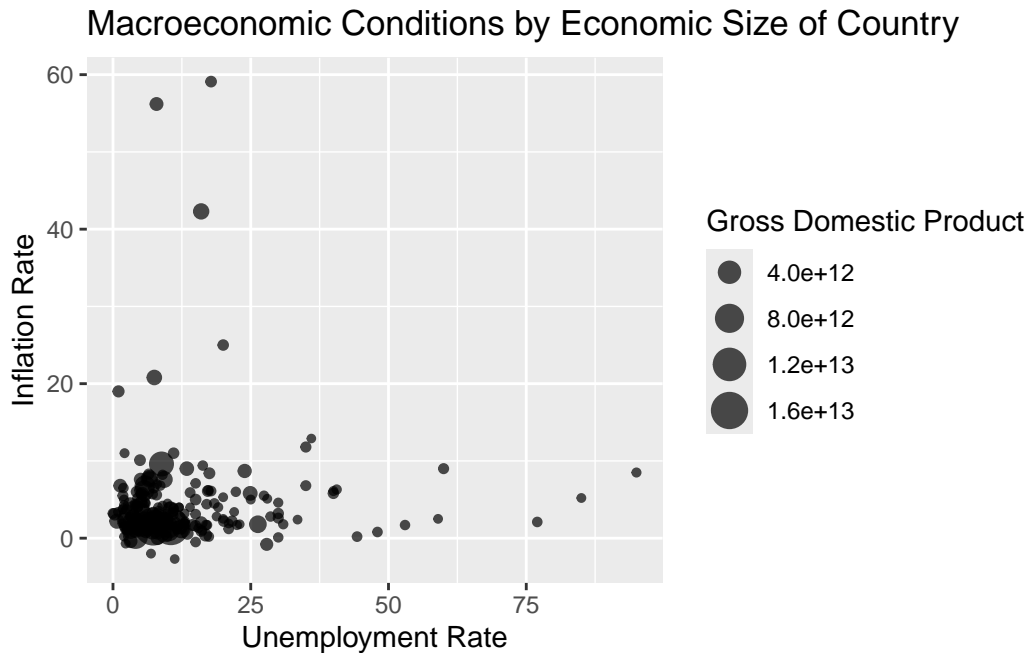
This plot illustrates the strong inverse relationship between the infant mortality rate and life expectancy of a country, incorporating health spending. Countries that invest more heavily in healthcare tend to have lower infant mortality and longer lifespans. This reflects real world

access to medical services, emphasizing how public policy choices have a direct influence on population health outcomes.



GDP, Unemployment, Inflation

This plot explores macroeconomic stability by examining unemployment rates and inflation, with GDP representing the overall economic scale of a country. Larger economies may better absorb any economic shocks, while smaller economies often face greater volatility. High unemployment combined with high inflation can significantly reduce quality of life for a population and limit policy options.



FAIR and CARE Principles

FAIR

findable: The data is easy to find. It is taken from the 2014 CIA Factbook which can be found online. The data is also easily accessible through the dcData package on RStudio.

accessible: The datasets are open, free, and universally implementable.

interoperable: The data uses a formal, accessible, shared, and broadly applicable language for knowledge representation. It provides a common understanding of all attributes and countries represented in the data.

reusable: The data is thoroughly described with multiple relevant and accurate attributes. CountryData contains 76 attributes for each country, and CountryCentroids contains 4 attributes.

CARE

One challenge we had with determining whether or not this data conformed to the CARE principles was assessing whether or not the datasets ignored power differentials and historical contexts related to Indigenous Peoples.

collective benefit: This data includes all countries and supports use and reuse by Indigenous nations. It enriches potential governance and engagement of citizens as it provides all nations and communities with a clear view of each country's attributes. The data benefits all communities in an equitable manner.

authority to control: The data is available and accessible to Indigenous Peoples. This way, the datasets support Indigenous governance.

responsibility: When using this data, we ensured that any representation of data from Indigenous nations were respectful of the dignity of all Indigenous nations and communities.

ethics: The datasets provide an accurate and fair view of all countries represented. This minimizes harm and maximizes benefit as it does not stigmatize Indigenous communities and represents all communities fairly.

Conclusion

Through multivariate visualizations and descriptive statistics, this exploratory data analysis reveals clear and consistent relationships among attributes of countries. The results suggest that investments in healthcare and education are strongly associated with improved life expectancy, reduced infant mortality, and greater countrywide stability. Macroeconomic patterns such as unemployment and inflation vary with the economic scale of a country. Our findings do not suggest complete causation, but they generate hypotheses for potential confirmatory data analysis. The data analysis emphasizes the value of looking into multiple variables when interpreting nationwide outcomes and establishes the role of data visualization as a powerful tool for understanding complex systems.

Code Appendix

```
# Load all necessary packages
library(tidyverse)
library(ggplot2)
library(ggpubr)
library(jpeg)
library(dcData)
library(kableExtra)
library(rvest)
library(readxl)
library(tibble)
library(dcData)

# Upload world map image
```

```

img <- readJPEG("/Users/gracerovinsky/Downloads/WorldMap.jpg")

# Select large countries
countries <- CountryData %>%
  select(c(country, birth, pop)) %>%
  filter(pop > 100000)

locations <- CountryCentroids %>%
  select(c(name, long, lat)) %>%
  rename(country = name)

dataPolished <- left_join(
  x = countries,
  y = locations,
  by = join_by(country == country)
)

# Create Population vs. Birth Rate visualization
birth_rate <- ggplot(
  data = dataPolished,
  mapping = aes(
    x = long,
    y = lat,
    size = pop/1000000,
    color = birth
  )) +
  background_image(img) +
  geom_point() +
  coord_cartesian(xlim = c(-180,180), ylim = c(-90,90)) +
  scale_color_gradient(low = 'red', high = 'blue') +
  labs(
    x = "Longitude",
    y = "Latitude",
    title = "Map of Population and Birth Rate",
    color = "Births per 1000 People",
  ) +
  guides(size = 'none')

# Read Country By Continent file
continents <- read_excel("/Users/gracerovinsky/Downloads/CountryByContinent.xlsx") %>%
  add_row(Brazil = "Brazil", `South America` = "South America", .before = 1) %>%
  rename(

```

```

    "country" = "Brazil",
    "continent" = "South America"
  )

countries <- CountryData

# Tidy country data
polished <- countries %>%
  select(c(1:6),12,16,18,26) %>%
  left_join(continents, by = "country") %>%
  drop_na() %>%
  rename("Continent" = "continent") %>%
  group_by(Continent) %>%
  summarise(
    Area_mean = mean(area),
    Pop_mean = mean(pop),
    Growth_mean = mean(growth),
    Birth_mean = mean(birth),
    Death_mean = mean(death),
    Health_mean = mean(health),
    Obesity_mean = mean(obesity),
    Educ_mean = mean(educ),
    Unemployment_mean = mean(unemployment)
  )

country_by_continent <- polished %>%
  kable( # Make table
    caption = "Country Data Summary Statistics, Grouped By Continent", # Add title
    align = c("l", rep("c", 6)) # Control the text alignment in each column
  ) %>%
  kable_classic(font_size = 16, lightable_options = "striped")

# Select key variables for summary
summary_stats <- CountryData %>%
  select(area, pop, growth, birth, death, health, obesity, educ, unemployment, life, infant,

# Create summary table for data
summary_tbl <- CountryData %>%
  select(
    area, pop, growth, birth, death,
    health, obesity, educ, unemployment, life

```

```

) %>%
psych::describe()

# Create Health Spending, Infant Mortality, and Life Expectancy visualization
health_visualization <- CountryData %>%
  drop_na(infant, life, health) %>% # Drop missing values for relevant variables
  ggplot(aes(x = infant, y = life)) +
  geom_point(aes(size = health), alpha = 0.7) +
  labs(
    title = "Health Spending and Population Health Outcomes",
    x = "Infant Mortality Rate",
    y = "Life Expectancy (years)",
    size = "Health Spending")

# Create GDP, Inflation, and Unemployment visualization
gdp_visualization <- CountryData %>%
  drop_na(GDP, unemployment, inflation) %>% # Drop missing values for relevant variables
  ggplot(aes(x = unemployment, y = inflation)) +
  geom_point(aes(size = GDP), alpha = 0.7) +
  labs(
    title = "Macroeconomic Conditions by Economic Size of Country",
    x = "Unemployment Rate",
    y = "Inflation Rate",
    size = "Gross Domestic Product"
  )

```