Inequality Evaluation

Hugo Jal

2024-06-01

This project is an assignment part of the Data Analysis with R course available at Udacity. The aim of this project is to evaluate the progression of the Gini Coefficient - a measure of inequality - across 197 countries from 1800 to 2050 (i.e., predictions until 2050). The data was extracted from Gapminder.

FREE DATA FROM WORLD BANK VIA GAPMINDER.ORG, CC-BY LICENSE

```
pf <- read.csv('gini.csv')
```

```
library(dplyr)
```

```
##
## Adjuntando el paquete: 'dplyr'

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

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

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.4.1

## Warning: package 'lubridate' was built under R version 4.4.1

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.5
## v ggplot2   3.5.1     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1

## -- 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 errors
```

```r
pf_long <- pf %>%
  pivot_longer(
    cols = starts_with("X"),
    names_to = "Year",
    names_prefix = "X",
    values_to = "Gini"
  )
```

```r
countries <- unique(pf_long$country)

# Create groups
countries_europe <- c("Albania", "Andorra", "Armenia", "Austria", "Azerbaijan", "Belarus", "Belgium", "
countries_africa <- c("Algeria", "Angola", "Benin", "Botswana", "Burkina Faso", "Burundi", "Cape Verde"
countries_asia <- c("Afghanistan", "Bahrain", "Bangladesh", "Bhutan", "Brunei", "Cambodia", "China", "T
countries_north_america <- c("Antigua and Barbuda", "Bahamas", "Barbados", "Belize", "Canada", "Costa R
countries_oceania <- c("Australia", "Fiji", "Kiribati", "Marshall Islands", "Micronesia, Fed. Sts.", "N
countries_south_america <- c("Argentina", "Bolivia", "Brazil", "Chile", "Colombia", "Ecuador", "Guyana"
```

```r
library(dplyr)

pf_long <- pf_long %>%
  mutate(
    continent = case_when(
      country %in% countries_europe ~ "Europe",
      country %in% countries_asia ~ "Asia",
      country %in% countries_africa ~ "Africa",
      country %in% countries_north_america ~ "North America",
      country %in% countries_south_america ~ "South America",
      country %in% countries_oceania ~ "Oceania",
      TRUE ~ "Other"  # Handle any countries not categorized
    )
  )
```

Gini Coefficient by Country To start with, my main focus is the present Gini coefficient in 2024 for each of the 197 nations. Because there are numerous countries, the chart displays them organized by continent based on color. Moreover, the plot is interactive such that hovering your mouse over any point will display the country and Gini coefficient.

```r
library(tidyverse)

# Filter data to 2024.
pf_2024 <- pf_long %>%
  filter(pf_long$Year == 2024)
```

```r
library(plotly)
```

```
## Warning: package 'plotly' was built under R version 4.4.1
```

```
##
## Adjuntando el paquete: 'plotly'

## The following object is masked from 'package:ggplot2':
##
##     last_plot

## The following object is masked from 'package:stats':
##
##     filter

## The following object is masked from 'package:graphics':
##
##     layout
```

```
p1 <-
  ggplot(data = pf_2024, aes(x = factor(country), y = Gini, , color = continent, text = paste("Country:
  geom_point(size = 3) +
  theme_minimal() +
  labs(title = "Gini Coefficients by Country and Continent (2024)",
       x = "Continent",
       y = "Gini Coefficient") +
  theme(legend.position = "none",
        axis.text.x = element_text(angle = 90, hjust = 1))

interactive_plot <- ggplotly(p1, tooltip = "text")

interactive_plot
```

```
## PhantomJS not found. You can install it with webshot::install_phantomjs(). If it is installed, pleas
```

Evolution of Mean Gini Coefficient by Continent

```
library(dplyr)

mean_by_continent <- pf_long %>%
  group_by(continent, Year) %>%
  summarise(mean_gini = mean(Gini, na.rm = TRUE))
```

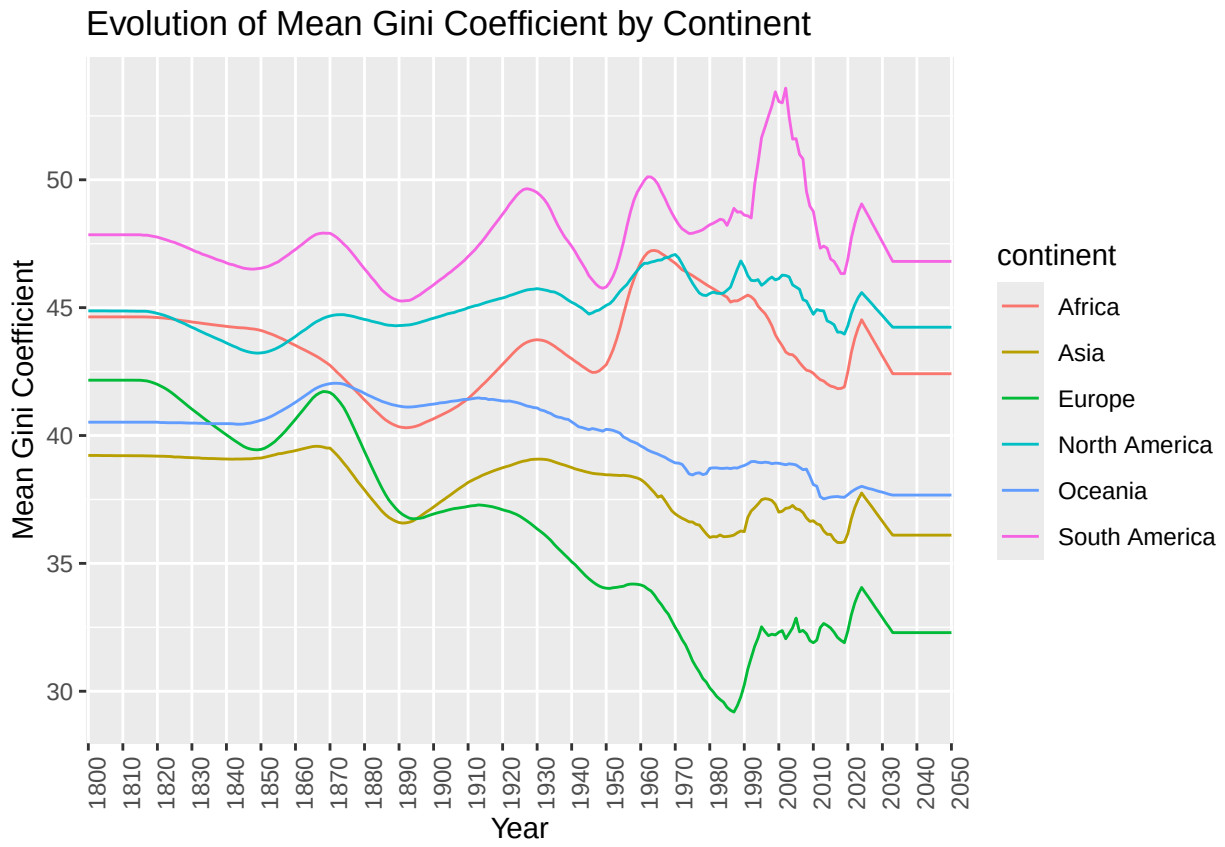
```
## `summarise()` has grouped output by 'continent'. You can override using the
## `.groups` argument.
```

```
p2 <-
  ggplot(mean_by_continent, aes(x = Year, y = mean_gini,
                                color = continent, group = continent)) +
  geom_line() +
  scale_x_discrete(breaks = seq(1800, 2050, 10)) +
  labs(x = 'Year', y = 'Mean Gini Coefficient', title = 'Evolution of Mean Gini Coefficient by Continen
  theme(axis.text.x = element_text(angle = 90))

ggsave(plot = p2, 'Evolution_of_Mean_Gini_Coefficient_by_Continent.png')
```

```
## Saving 6.5 x 4.5 in image
```

```
print(p2)
```




```
p3 <- p2 + facet_wrap(~continent) # To show each continent separate

ggsave(plot = p3, 'Evolution_of_Mean_Gini_Coefficient_by_Continent_separated.png')
```

```
## Saving 6.5 x 4.5 in image
```

South America consistently has the highest inequality, with significant fluctuations over time. Europe shows a notable decline in inequality, especially from the mid-20th century onward, reaching the lowest levels among all continents by 2050. Africa and North America maintain relatively stable mid-range coefficients throughout the period. Asia experiences an increase in inequality from the mid-20th century, while Oceania shows some variability but remains in the middle range.

Another interesting insight is that in the early 2020s, there was a notable increase in inequality across all continents, probably due to the COVID-19 pandemic and how it hindered the market at that time.

Normality & Summary Statistics for 2024

```
pf_africa <- subset(pf_2024, pf_2024$continent == 'Africa')
pf_asia <- subset(pf_2024, pf_2024$continent == 'Asia')
pf_europe <- subset(pf_2024, pf_2024$continent == 'Europe')
pf_oceania <- subset(pf_2024, pf_2024$continent == 'Oceania')
pf_na <- subset(pf_2024, pf_2024$continent == 'North America')
pf_sa <- subset(pf_2024, pf_2024$continent == 'South America')
```

```r
shapiro.test(pf_africa$Gini)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  pf_africa$Gini
## W = 0.96186, p-value = 0.08345
```

```r
shapiro.test(pf_asia$Gini)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  pf_asia$Gini
## W = 0.962, p-value = 0.1545
```

```r
shapiro.test(pf_europe$Gini)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  pf_europe$Gini
## W = 0.97249, p-value = 0.2914
```

```r
shapiro.test(pf_oceania$Gini)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  pf_oceania$Gini
## W = 0.95494, p-value = 0.6397
```

```r
shapiro.test(pf_na$Gini)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  pf_na$Gini
## W = 0.93288, p-value = 0.1262
```

```r
shapiro.test(pf_sa$Gini)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  pf_sa$Gini
## W = 0.8645, p-value = 0.05568
```

```r
# All continents are normally distributed for 2024 (i.e., p value > 0.05)
```

```r
library(dplyr)

summary_stats <- pf_2024 %>%
  group_by(continent) %>%
  summarise(
    Mean_Gini = mean(Gini, na.rm = TRUE),
    Median_Gini = median(Gini, na.rm = TRUE),
    SD_Gini = sd(Gini, na.rm = TRUE)
  )
print(summary_stats)
```

```
## # A tibble: 6 x 4
##   continent     Mean_Gini Median_Gini SD_Gini
##   <chr>             <dbl>       <dbl>   <dbl>
## 1 Africa             44.5        43.5    8.53
## 2 Asia               37.8        37.3    4.79
## 3 Europe             34.1        34.3    4.58
## 4 North America      45.6        45.4    6.77
## 5 Oceania            38.0        37.6    2.74
## 6 South America      49.1        46.8    5.93
```

This table shows quite significant standard deviations for 2024 in all continents but Oceania. Europe and Asia still have "acceptable" standard deviations; however, the rest of the continents show signs of great inequality amongst the intra-continental group.