COVID Data Assignment

Stark, Joop

2025-11-25

Introduction

This analysis is to see if a countries wealth effected the survival rate of its people. There are many proxies for a countries' wealth this study will compare the survival rating to a countries' "Gross Domestic Product", "Gross Domestic Product per Capita", and an informal Price Point Parity, the "Big Mac Index".

The analysis fortunately shows that these three countries' wealth proxies play a minimal role on there people's chances of surviving COVID.

Libraries

Tidyverse was used in this analysis, but should have used a library to convert countries into uniform codes.

```
library(tidyverse)
```

Import Data

COVID data Johns Hopkins' Center for Systems Science and Engineering.

```
url_in <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/refs/heads/master/csse_covid_19_da

file_names <- c(
  "time_series_covid19_confirmed_US.csv",
  "time_series_covid19_confirmed_global.csv",
  "time_series_covid19_deaths_US.csv",
  "time_series_covid19_deaths_global.csv"
)

url <- str_c(url_in, file_names)

confirmed_us <- read.csv(url[1])
confirmed_global <- read.csv(url[2])
deaths_us <- read.csv(url[3])
deaths_global <- read.csv(url[4])
```

The economic and population data was downloaded as a zip from Kaggle, it can be accessed from the links provided or the data directory from the github link. Github: https://github.com/JoopStark/ds_field_covid_assignment

```r
# https://www.kaggle.com/datasets/alejopaullier/-gdp-by-country-1999-2022
# https://www.kaggle.com/datasets/vittoriogiatti/bigmacprice?resource=download

world_gdp <- read.csv("data/gdp.csv")
big_mac_price <- read.csv("data/BigmacPrice.csv")
```

```r
# https://www.kaggle.com/datasets/tanuprabhu/population-by-country-2020

population<- read.csv("data/population_by_country_2020.csv")
```

Tidy

Tidy COVID

The analysis only needed yearly data so cases and deaths were summed for each year. Then deaths and case were joined onto a combined COVID table. This table will be combined with the economic data, so some country names needed to be converted to match other data.

```r
# Global data
confirmed_global_tidy <- confirmed_global |>
  select(!c(Province.State ,Lat, Long)) |>
  pivot_longer(cols = !c(Country.Region),
               names_to = "date",
               values_to = "cases") |>
  mutate(year = year(mdy(substr(date, 2, nchar(date))))) |>
  group_by(Country.Region, year) |>
  summarise(cases = sum(cases))

deaths_global_tidy <- deaths_global |>
  select(!c(Province.State ,Lat, Long)) |>
  pivot_longer(cols = !c(Country.Region),
               names_to = "date",
               values_to = "deaths") |>
  mutate(year = year(mdy(substr(date, 2, nchar(date))))) |>
  group_by(Country.Region, year) |>
  summarise(deaths = sum(deaths))

#join table next

join_global <- confirmed_global_tidy |>
  full_join(deaths_global_tidy, by=c("Country.Region", "year")) |>
  mutate(Country.Region = if_else(Country.Region == "Taiwan*", "Taiwan", Country.Region),
         Country.Region = if_else(Country.Region == "Czechia", "Czech Republic", Country.Region),
         Country.Region = if_else(Country.Region == "Korea, South", "South Korea", Country.Region),
         Country.Region = if_else(Country.Region == "Congo (Brazzaville)", "Congo, Republic of", Country
         Country.Region = if_else(Country.Region == "Congo (Kinshasa)", "Congo, Democratic Republic of"
         )

# I thought I need this but did not
# US data
# confirmed_us_tidy <- confirmed_us |>
```

```
#   select(!c(UID, iso3:Combined_Key)) |>
#   pivot_longer(cols = !c(iso2),
#                names_to = "date",
#                values_to = "cases") |>
#   mutate(year = year(mdy(substr(date, 2, nchar(date))))) |>
#   group_by(year) |>
#   summarise(cases = sum(cases))
#
# deaths_us_tidy <- deaths_us |>
#   select(!c(UID, iso3:Combined_Key)) |>
#   pivot_longer(cols = !c(iso2, Population),
#                names_to = "date",
#                values_to = "deaths") |>
#   mutate(year = year(mdy(substr(date, 2, nchar(date))))) |>
#   group_by(year) |>
#   summarise(deaths = sum(deaths), population = sum(Population))
#
# joined_us <- confirmed_us_tidy |>
#   left_join(deaths_us_tidy, by="year")
#
```

Tidy Economic

```
world_gdp_tidy <- world_gdp |>
  select(Country, X2019:X2022) |>
  mutate(across(X2019:X2022, parse_number)) |>
  pivot_longer(cols = !c(Country),
               names_to = "year",
               values_to = "gdp") |>
  mutate(year = year(make_date(substr(year, 2, nchar(year))))) |>
  #lag gdp
  arrange(Country, year) |>
  group_by(Country) |>
  mutate(previous_gdp = lag(gdp)) |>
  ungroup() |>
  filter(year >= 2020 & year <= 2023) |>
  mutate( Country = case_when(
    Country == "Taiwan Province of China" ~ "Taiwan",
    Country == "United States" ~ "US",
    Country == "Korea" ~ "South Korea",
    Country == "Slovak Republic" ~ "Slovakia",
    Country == "Afghanistan, Rep. of." ~ "Afghanistan",
    Country == "Bahamas, The" ~ "Bahamas",
    Country == "Burma" ~ "Myanmar",
    Country == "Gambia, The" ~ "Gambia",
    Country == "Iran, Islamic Republic of" ~ "Iran",
    Country == "Kyrgyz Republic" ~ "Kyrgyzstan",
    Country == "Lao People's Democratic Republic" ~ "Laos",
    Country == "Macedonia" ~ "North Macedonia",
    Country == "Syrian Arab Republic" ~ "Syrian",
    Country == "Timor-Leste, Dem. Rep. of" ~ "Timor-Leste",
    Country == "Yemen, Republic of" ~ "Yemen",
```

```
    Country == "Macedonia, Former Yugoslav Republic of" ~ "North Macedonia",
    TRUE ~ Country
  ))


big_mac_price_tidy <- big_mac_price |>
  select(date, name, dollar_price) |>
  mutate(date = ymd(date)) |>
  rename(big_mac_usd = dollar_price) |>
  group_by(year = year(date), name) |>
  summarize(average_big_mac_usd= mean(big_mac_usd))|>
  #lag price one year
  arrange(name, year) |>
  group_by(name) |>
  mutate(previous_big_mac_usd= lag(average_big_mac_usd)) |>
  ungroup() |>
  filter(year >= 2020 & year <= 2023) |>
  filter(!name %in% c("Hong Kong", "Euro area")) |>
  mutate(name = if_else(name == "United States", "US", name),
         name = if_else(name == "Britain", "United Kingdom", name))
```

Tidy Population

```
population_tidy <- population |>
  select(Country..or.dependency., Population..2020.) |>
  rename(country = Country..or.dependency., population = Population..2020.) |>
  mutate(
    country = case_when(
      country == "United States" ~ "US",
      country == "Czech Republic (Czechia)" ~ "Czech Republic",
      country == "Burma" ~ "Myanmar",
      country == "DR Congo" ~ "Congo, Democratic Republic of",
      country == "Congo" ~ "Congo, Republic of",
      TRUE ~ country
    )
  )
```

Main Table Created

```
global_covid_economic <- big_mac_price_tidy |>
  full_join(world_gdp_tidy, join_by(name == Country, year)) |>
  full_join(join_global, join_by(name == Country.Region, year)) |>
  full_join(population_tidy, join_by(name == country)) |>
  rename(country = name)
```

Transformation

Columns added are `death_rate`, `survival_rate`, `survival_scaled_mean`, `survival_zscore`, and `gdp_per_cap`. The Survival Mean and Z-score where not used. GDP was in billions of dollars so to get the GDP per capita in need to multiply GDP by 1 billion before dividing by the population.

```
global_covid_economic_transform <- global_covid_economic |>
  mutate(death_rate = deaths/cases,
         survival_rate = 1 - death_rate,
         survival_scaled_mean = survival_rate - mean(survival_rate),
         survival_zscore = (survival_rate - mean(survival_rate)) / sd(survival_rate),
         gdp_per_cap = (gdp * 1000000000)/ population
         )
```

Outliers

There was an interest in investigating the countries that have lower survival rates, but removing them did not effect the r-squared value enough to warrant an investigation.

```
global_covid_economic_transform_wo_outliers <- global_covid_economic_transform |>
  filter(survival_rate > .96)
```

Models

```
big_mac_mod <- lm(survival_rate ~ average_big_mac_usd, data=global_covid_economic_transform, na.action =
gdp_mod <- lm(survival_rate ~ gdp, data=global_covid_economic_transform, na.action = na.exclude)
gdp_per_cap_mod <- lm(survival_rate ~ gdp_per_cap, data= global_covid_economic_transform, na.action = na

gdp_per_cap_wo_mod <- lm(survival_rate ~ gdp_per_cap, data= global_covid_economic_transform_wo_outliers)
```

R^2

The low R^2 values show that the models do not closely represent the data.

```
summary(big_mac_mod)$r.squared
```

```
## [1] 0.014516
```

```
summary(gdp_mod)$r.squared
```

```
## [1] 0.0006645189
```

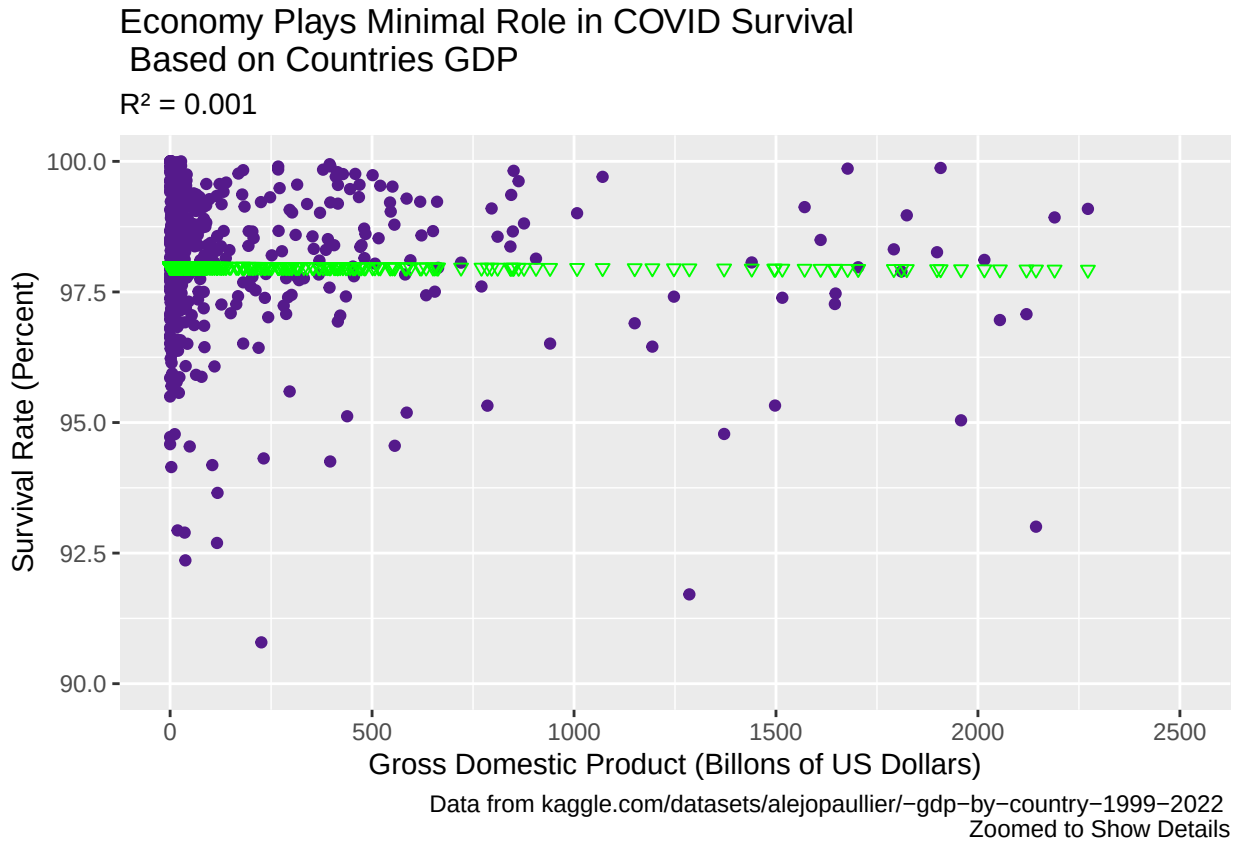
```
summary(gdp_per_cap_mod)$r.squared
```

```
## [1] 0.02198275
```

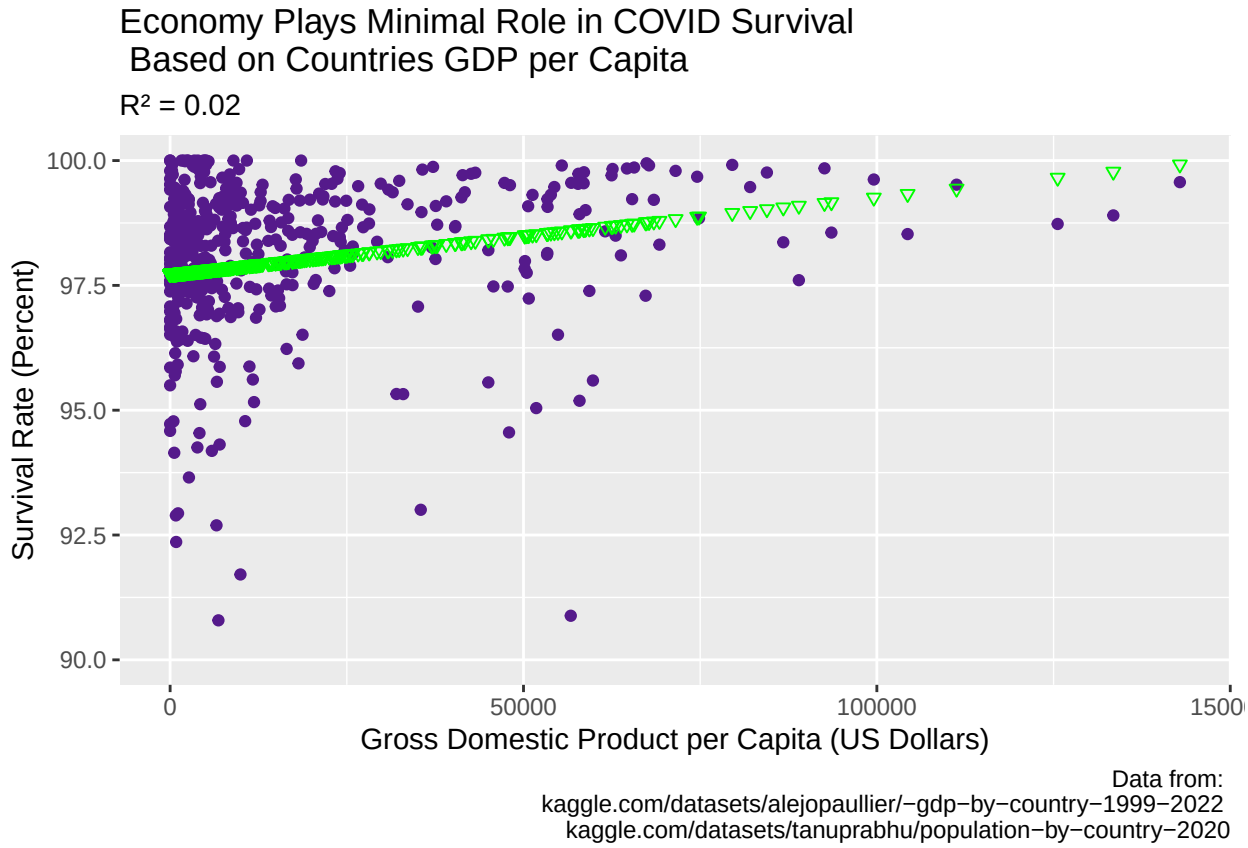
```
global_covid_economic_transform <-  global_covid_economic_transform |>
  mutate(
    big_mac_pred = predict(big_mac_mod),
    gdp_pred = predict(gdp_mod),
    gdp_per_cap_pred = predict(gdp_per_cap_mod)
  )
```

Graphs

```
global_covid_economic_transform |>
  ggplot() +
  geom_point(aes(x = gdp, y = survival_rate * 100), color="purple4") +
  geom_point(aes(x = gdp, y = gdp_pred * 100), color="green", shape = 6) +
  xlim(0, 2500) +
  ylim(90, 100) +
  labs(
    x = "Gross Domestic Product (Billons of US Dollars)",
    y = "Survival Rate (Percent)",
    title = "Economy Plays Minimal Role in COVID Survival \n Based on Countries GDP",
    subtitle = "R² = 0.001",
    caption = "Data from kaggle.com/datasets/alejopaullier/-gdp-by-country-1999-2022 \n Zoomed to Show D
  )
```

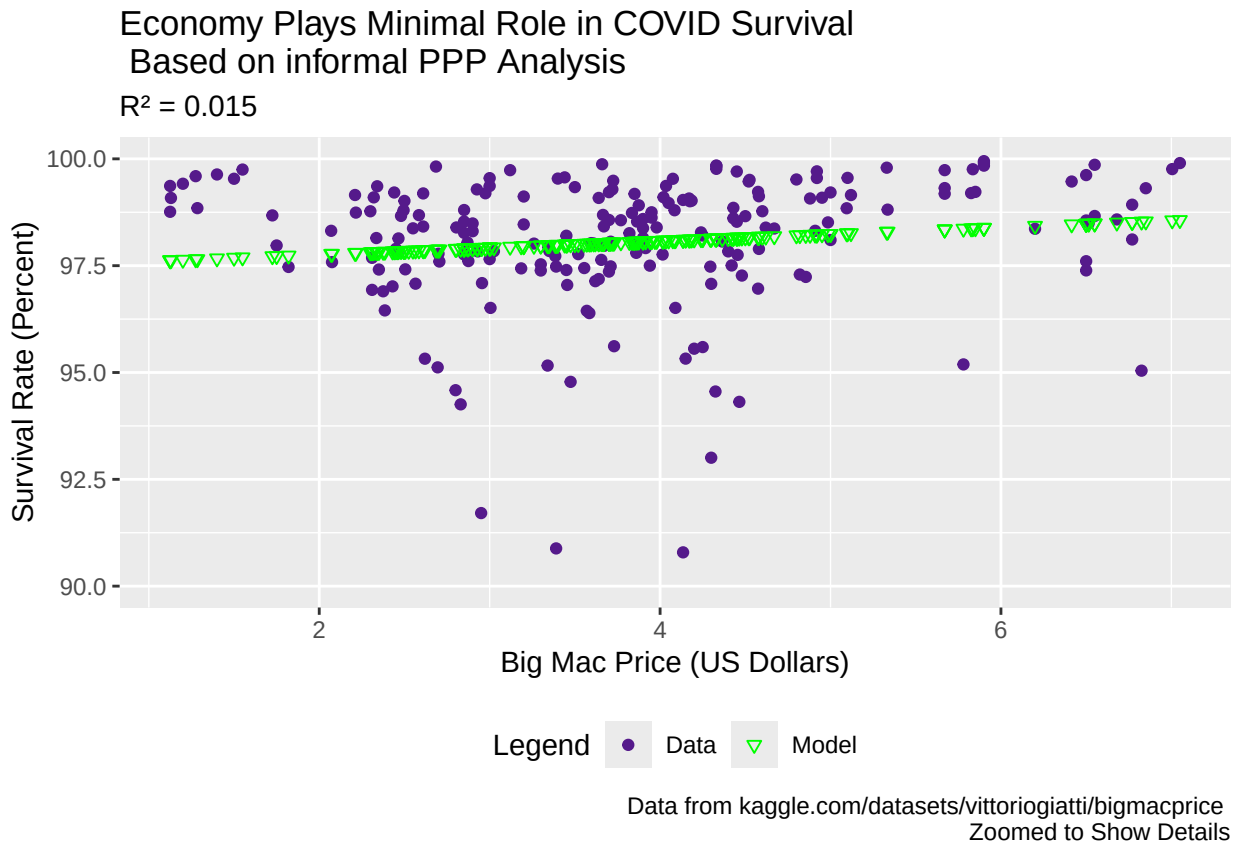


```
global_covid_economic_transform |>
  ggplot() +
  geom_point(aes(x = gdp_per_cap, y = survival_rate * 100), color="purple4") +
  geom_point(aes(x = gdp_per_cap, y = gdp_per_cap_pred * 100), color="green", shape = 6) +
  ylim(90, 100) +
  labs(
    x = "Gross Domestic Product per Capita (US Dollars)",
    y = "Survival Rate (Percent)",
    title = "Economy Plays Minimal Role in COVID Survival \n Based on Countries GDP per Capita",
    subtitle = "R² = 0.02",
    caption = "Data from: \n kaggle.com/datasets/alejopaullier/-gdp-by-country-1999-2022 \n kaggle.com/
  )
```




```
global_covid_economic_transform |>
  ggplot() +
  geom_point(aes(x = average_big_mac_usd, y = survival_rate * 100, color = "Data") ) +
  geom_point(aes(x = average_big_mac_usd, y = big_mac_pred * 100, color = "Model"), shape = 6) +
  ylim(90, 100) +
  scale_color_manual(
    name = "Legend",
    values = c("Data" = "purple4", "Model" = "green")
  )+
  theme(legend.position = "bottom")+
  labs(
    x = "Big Mac Price (US Dollars)",
```

```
    y = "Survival Rate (Percent)",
    title = "Economy Plays Minimal Role in COVID Survival \n Based on informal PPP Analysis",
    subtitle = "R² = 0.015",
    caption = "Data from kaggle.com/datasets/vittoriogiatti/bigmacprice \n Zoomed to Show Details"
)
```



Conclusion

After an analysis of data to include graphs. It appears that the wealth of a country does not have a huge effect on its people's chance of surviving COVID at least by itself.

Limitations of analysis

- Biased data:
  - Opinions on COVID effected what was reported as COVID. Sometimes COVID related deaths were reported as other illness, and sometimes unrelated deaths were attributed to COVID because the individual had COVID.
  - Wealth metrics and population can be effect by the beliefs of the organizations taking them.
  - Not all countries have the resources to properly diagnose or report COVID
- This analysis does not account for:

- different version of COVID
- different cultures belief that effect care and spread
- effects of support of other countries, or NGOs
- effects of genetics of groups of people
- data from different resources may consider lands part of different countries