Socioeconomic Factors: Evaluating the Influence of Resources and Healthcare Access on COVID-19 Mortality

Jansen Smith

2023-08-22

Purpose

The purpose of this analysis is to assess the impact of resources and access to health care on COVID-19 mortality. This is achieved by investigating the relationships between a country's per capita income, which serves as a proxy for available resources, and life expectancy, which acts as a proxy for general access to health care. By examining these relationships and their correlation with normalized deaths per case due to COVID-19, we aim to uncover insights into how socioeconomic factors may influence COVID-19 outcomes on a global scale.

Data Source and Description

The primary data sources for this analysis are COVID-19 cases and deaths data obtained from the John Hopkins University's GitHub repository. Additionally, we are using Hans Rosling's dataset from GapMinder, which provides information on income, life expectancy, population, and region for various countries.

Importing required libraries

The only libraries required to run the following analysis are tidyverse and lubridate, and this code block is designed to automatically install tidyverse if the user has not already done so.

```
# Install the tidyverse package if not already installed
if (!requireNamespace("tidyverse", quietly = TRUE)) {
  install.packages("tidyverse")
}

# Load the tidyverse package
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- 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
# Load the lubridate package for parsing dates
library(lubridate)
```

COVID-19 data import

Begin by reading in cases and deaths data from two main csv files, over global scope.

```r
# These two sources were provided by the professor.
url = "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19
filenames = c("time_series_covid19_confirmed_global.csv",
              "time_series_covid19_deaths_global.csv")
urls = str_c(url, filenames)

# Read CSV files and suppress messages about column types
global_cases = read_csv(urls[1], show_col_types = FALSE)
global_deaths = read_csv(urls[2], show_col_types = FALSE)
```

Tidy global data

To facilitate analysis, the global COVID-19 cases and deaths data are tidied and transformed. This involves renaming columns, selecting relevant date columns, and pivoting the data to a longer format. By organizing the data in this manner, we create a structured dataset that is amenable to further manipulation and analysis.

```r
# Tidying the global cases, as demonstrated by the professor.
global_cases_tidy = global_cases %>%
  rename(Province_State = "Province/State", Country_Region = "Country/Region") %>%
  select("Province_State", "Country_Region", matches("^\\d{1,2}/\\d{1,2}/\\d{2}$")) %>%
  pivot_longer(cols = -c('Province_State',
                         'Country_Region'),
               names_to = "date",
               values_to = "cases") %>%
  mutate(date = as.Date(date, format = "%m/%d/%y"))

# Tidying the global deaths, as demonstrated by the professor.
global_deaths_tidy <- global_deaths %>%
  rename(Province_State = "Province/State", Country_Region = "Country/Region") %>%
  select("Province_State", "Country_Region", matches("^\\d{1,2}/\\d{1,2}/\\d{2}$")) %>%
  pivot_longer(cols = -c('Province_State', 'Country_Region'),
               names_to = "date",
               values_to = "deaths") %>%
  mutate(date = as.Date(date, format = "%m/%d/%y"))
```

Join datasets

To facilitate a comprehensive analysis that incorporates both COVID-19 cases and deaths data, the two tidied datasets, `global_cases_tidy` and `global_deaths_tidy`, are merged. This enables the exploration of cases and deaths as separate variables within the same dataset.

```r
# Merge global_cases_tidy and global_deaths_tidy to create a combined dataset
suppressMessages({ # Messages suppressed for aesthetic reasons. Code runs without error.
  global = global_cases_tidy %>%
    full_join(global_deaths_tidy) %>%
    filter(cases > 0) %>%
    unite("Combined_Key", c(Province_State, Country_Region),
          sep = ", ",
          na.rm = TRUE,
          remove = FALSE)
})
```

Global Lifespan & Income Per Capita Import

Importing Hans Rosling's custom dataset for The Joy of Stats, maintained by his organization GapMinder since his death in 2017. This dataset contains essential socio-economic indicators, including lifespan, income per capita, population, and region, for various countries. The dataset serves as a valuable resource to explore the relationships between these indicators and COVID-19's impact on populations.

```r
# This csv URL is a permalink that will remain reproducible.
global_stats_url =
  "https://raw.githubusercontent.com/JansenSmith/COVID-Analysis/main/JoyOfStats_data.csv"
global_stats = read_csv(global_stats_url, show_col_types = FALSE)
```

Data Preparation and Latest Year Summary

In this section, we prepare the data for further analysis by focusing on the latest available date in the Johns Hopkins dataset.

To gain insights into COVID-19 lethality, we aggregate the data at the country level. This aggregation involves determining the total cases and deaths for each country. Countries with zero cases are removed from the dataset to ensure the analysis is based on relevant data.

For consistency, we standardize the names of certain countries. For instance, "US" is replaced with "United States," and "Burma" is replaced with "Myanmar."

```r
# Calculate the latest date in the global dataset
latest_date = max(global$date)

# Find the year of the latest date
latest_year = year(latest_date)

# Filter global_cases_tidy to get the latest data
latest_data = global %>%
  filter(date == latest_date)

# Aggregate the data at the country level by summing cases and deaths
aggregated_data = latest_data %>%
  group_by(Country_Region) %>%
  summarize(cases = sum(cases), deaths = sum(deaths)) %>%
  filter(cases > 0) %>%  # Remove countries with no cases
  mutate(Country_Region = ifelse(Country_Region == "US",
                                 "United States", Country_Region)) %>%
```

```
  mutate(Country_Region = ifelse(Country_Region == "Burma",
                                 "Myanmar", Country_Region)) %>%
  mutate(Country_Region = ifelse(Country_Region == "Taiwan*",
                                 "Taiwan", Country_Region))


# Filter the global_stats dataframe to isolate the relevant year
global_stats_latest_year = global_stats %>%
  filter(latest_year == as.numeric(str_extract(`Year 1`, "\\d{4}"))) %>%
  rename(Life_Expectancy = "Life Expectancy")
```

Deaths per Case Calculation and Global Average

To gain insights into the impact of COVID-19, we are computing the deaths per case ratio for each country by dividing the total deaths by the corresponding total cases. Furthermore, we are establishing the global average deaths per case by aggregating the deaths and cases across all countries and then calculating the ratio. This metric serves as a baseline for normalization.

```
# Calculate deaths per case ratio for each country based on the latest data
global_ratio = aggregated_data %>%
  mutate(deaths_per_case = deaths / cases) %>%
  filter(!is.na(deaths_per_case))  # Remove NA values

# Calculate global average deaths per case ratio
global_avg_ratio = global_ratio %>%
  summarise(global_avg_deaths_per_case = sum(deaths) / sum(cases))
```

Normalization Strategy

Given the inherent challenges posed by extremely small values resulting from deaths per case calculations, we are implementing a normalization technique. By dividing each country's deaths per case ratio by the global average, we are creating a normalized deaths per case metric. This adjustment ensures that the analysis remains robust and interpretable across the entire dataset.

```
# Calculate normalized deaths per case metric
global_ratio_normalized = global_ratio %>%
  mutate(normalized_deaths_per_case =
           deaths_per_case / global_avg_ratio$global_avg_deaths_per_case)
```

Data Consolidation and Error Handling

In this section, we consolidate the relevant data from different sources and perform error handling to ensure the quality of our analysis.

We start by merging the normalized deaths per case ratios with socio-economic indicators to combine COVID-19 mortality metrics with socio-economic context.

We then identify and explicitly remove any countries with erroneous data or potential outlier status. The country "Yemen" is removed due to its explicit outlier status, which could skew the overall analysis.

Finally, we print a list of the removed countries to clearly document the excluded data points and ensure transparency in the analysis. These removed countries will be discussed further in the "Bias Assessment" section to address potential biases in the analysis.

```r
# Merge global_ratio_normalized with global_stats, filtering appropriately
combined_data = global_ratio_normalized %>%
  left_join(global_stats_latest_year,
            by = c("Country_Region" = "country (income_per_person_gdpperc.csv)")) %>%
  select(Country_Region, normalized_deaths_per_case,
         Income, Life_Expectancy, Population, Region)

# Identify and remove countries with erroneous data
combined_data_filtered = combined_data %>%
  filter(!is.na(Income) & !is.na(normalized_deaths_per_case) & !is.na(Life_Expectancy) &
           !is.na(Population) & !is.na(Region)) %>%
  filter(Country_Region != "Yemen") # Explicitly removed due to outlier status

# Determine removed countries
removed_countries =
  unique(combined_data$Country_Region[!combined_data$Country_Region %in%
                                        combined_data_filtered$Country_Region])

# Create a copy of combined_data that only contains the removed countries
removed_data = combined_data %>%
  filter(Country_Region %in% removed_countries)

# Calculate the total population in removed_data
removed_population = sum(removed_data$Population, na.rm = TRUE)

# Print the list of removed countries
cat("Removed countries were", paste(removed_countries, collapse = ", "), ".\n\n")
```

Removed countries were Andorra, Antarctica, Cabo Verde, Congo (Brazzaville), Congo (Kinshasa), Czechia, Diamond Princess, Dominica, Eswatini, Holy See, Korea, North, Korea, South, Kosovo, Kyrgyzstan, Laos, Liechtenstein, MS Zaandam, Marshall Islands, Micronesia, Monaco, Nauru, North Macedonia, Palau, Saint Kitts and Nevis, Saint Lucia, Saint Vincent and the Grenadines, San Marino, Slovakia, Summer Olympics 2020, Taiwan, Tuvalu, West Bank and Gaza, Winter Olympics 2022, Yemen .

```r
# Print the number of removed persons
if (is.na(removed_population)) {
  cat("Population data is missing for all removed countries.\n")
} else if (removed_population == 0) {
  cat("No removed countries with valid population data.\n")
} else {
  removed_population = format(removed_population, big.mark = ",")
  cat("At least", removed_population, "unique human beings have been disregarded in this analysis, due t
}
```

At least 35,290,600 unique human beings have been disregarded in this analysis, due to technical difficulties.

Predictive Model Construction & Training

In this phase, our focus turns to constructing predictive models to unravel the potential connections between Income and Life Expectancy with normalized deaths per case. Two distinct predictive models are developed: one utilizing Income as a predictor (mod_Income), and the other employing Life Expectancy ($mod_Life_Expectancy$). These models aim to capture patterns that could shed light on the relationship between socio-economic indicators and the severity of COVID-19 outcomes.

We initiate the process by building linear regression models using the lm() function. The models are trained on the combined_data dataset, establishing relationships between these two stand-ins for resources and health care vs COVID mortality.

```
# Create two predictive models, using both Income and Life Expectancy to
#  predict normalized_deaths_per_case
mod_Income = lm(normalized_deaths_per_case ~ Income,
                data = combined_data_filtered)
mod_Life_Expectancy = lm(normalized_deaths_per_case ~ Life_Expectancy,
                         data = combined_data_filtered)
```

Predictive Model Assessment

In this section, we evaluate the performance and significance of the predictive models constructed earlier. By analyzing the model summaries, we can gauge the strength of the relationships and evaluate the statistical significance of the predictors.

To interpret the models' effectiveness, we employ the summary() function on both mod_Income and mod_Life_Expectancy. This provides us with valuable information about the coefficients, p-values, and R-squared values, helping us understand the predictive power and significance of each model.

Please note that the specifics of this assessment may change, as new data is added to the datasets over time.

Predictive Model: Income vs. Normalized Deaths per Case

Model Summary The linear regression model assessing the impact of **Income** on **Normalized Deaths per Case** reveals valuable insights:

- The intercept term estimates the **average normalized deaths per case** when Income is zero. In our context, this value might not hold any practical meaning.
- The coefficient of **Income** (-2.224e-05) represents the change in normalized deaths per case associated with a one-unit change in Income. The p-value (5.62e-08) suggests that the coefficient is statistically significant, indicating that Income is likely to have a significant impact on COVID-19 mortality.
- The **Adjusted R-squared** value (0.159) signifies that around 15.9% of the variability in normalized deaths per case can be explained by Income in the model.
- The **F-statistic** (32.4) along with the low p-value (5.62e-08) suggests that the model as a whole is statistically significant.

This model suggests that countries with higher Income tend to have lower normalized deaths per case. However, it's important to note that the effect size is small, and other factors not included in the model may also play a role.

Predictive Model: Life Expectancy vs. Normalized Deaths per Case

Model Summary The linear regression model examining the influence of **Life Expectancy** on **Normalized Deaths per Case** yields the following insights:

- The intercept term (5.01292) represents the estimated average normalized deaths per case when Life Expectancy is zero, which lacks practical interpretation.
- The coefficient of **Life Expectancy** (-0.04886) indicates that for every one-unit increase in Life Expectancy, the normalized deaths per case decrease by approximately 0.04886. The p-value (0.000188) suggests that the coefficient is statistically significant, implying that Life Expectancy likely has a significant impact on COVID-19 outcomes.

- The **Adjusted R-squared** value (0.07572) signifies that around 7.6% of the variability in normalized deaths per case can be explained by Life Expectancy in the model.
- The **F-statistic** (14.6) along with the low p-value (0.000188) indicates that the model as a whole is statistically significant.

This model suggests that countries with higher Life Expectancy tend to experience lower normalized deaths per case. However, similar to the previous model, the effect size is modest, and other unaccounted factors can contribute to COVID-19 outcomes.

These models provide us with insights into how Income and Life Expectancy are related to COVID-19 mortality. However, it's crucial to acknowledge that correlation does not imply causation. Other factors not included in the models may be influencing the observed relationships.

```
# Assess the predictive models for income and life expectancy's impact
#  on normalized deaths per case
summary(mod_Income)
```

```
##
## Call:
## lm(formula = normalized_deaths_per_case ~ Income, data = combined_data_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7576 -0.6493 -0.2026  0.4023  5.9913
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.840e+00  1.135e-01  16.216  < 2e-16 ***
## Income      -2.224e-05  3.908e-06  -5.692 5.62e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.059 on 165 degrees of freedom
## Multiple R-squared:  0.1641, Adjusted R-squared:  0.159
## F-statistic:  32.4 on 1 and 165 DF,  p-value: 5.62e-08
```

```
summary(mod_Life_Expectancy)
```

```
##
## Call:
## lm(formula = normalized_deaths_per_case ~ Life_Expectancy, data = combined_data_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8360 -0.6716 -0.2804  0.4511  6.2176
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.01292    0.95118   5.270 4.21e-07 ***
## Life_Expectancy -0.04886    0.01279  -3.821 0.000188 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.11 on 165 degrees of freedom
## Multiple R-squared:  0.08129,    Adjusted R-squared:  0.07572
## F-statistic:  14.6 on 1 and 165 DF,  p-value: 0.000188
```

Dataset Augmentation Based On Predictive Models

To deepen our analysis, we enhance the combined_data dataset by incorporating the predictive model outcomes. This augmentation involves the creation of new variables, Income_pred and Life_Expectancy_pred, which capture the predicted values from the respective models. These predictions will be reflected in the final plots, denoted in red.

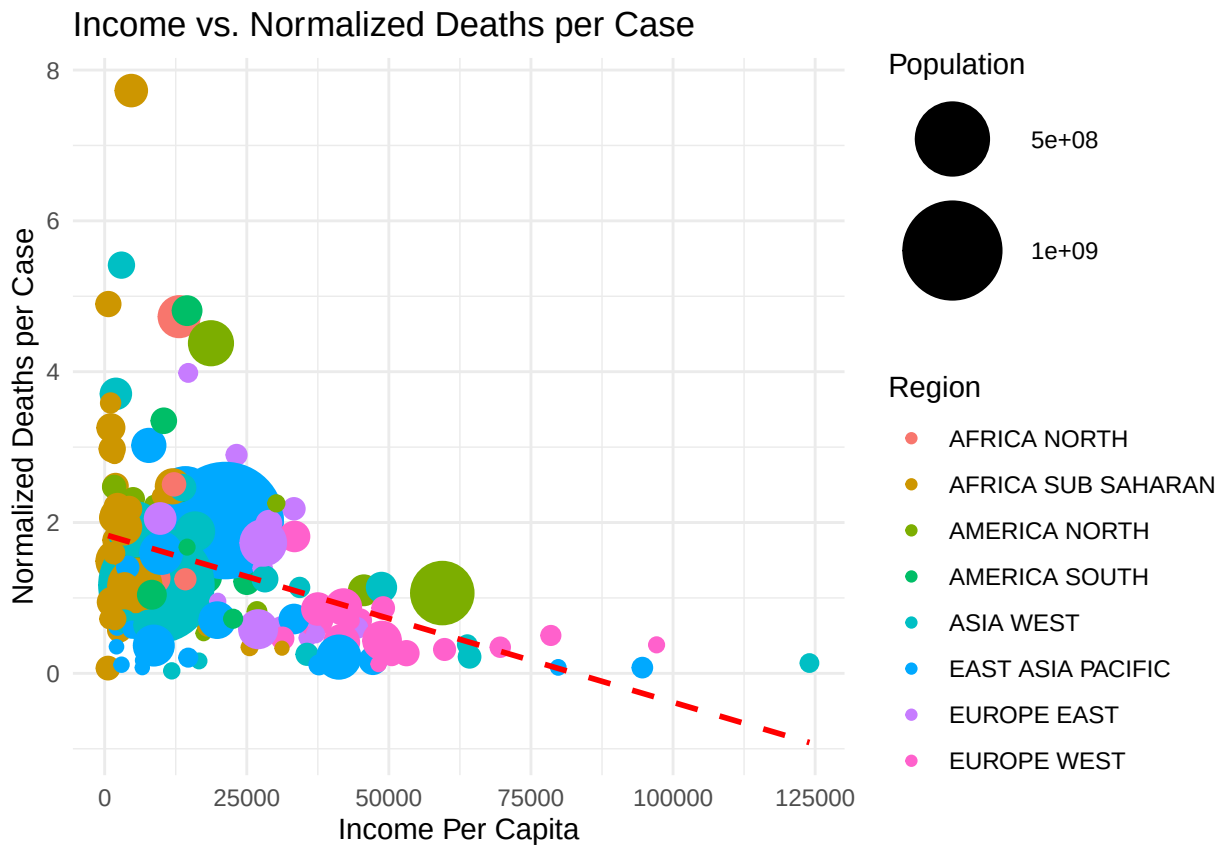
```
# Enhance the combined_data dataset by incorporating predictive model outcomes
combined_data_predictive = combined_data_filtered %>%
  mutate(Income_pred = predict(mod_Income),
         Life_Expectancy_pred = predict(mod_Life_Expectancy)) %>%
  select(1:Income, Income_pred, Life_Expectancy, Life_Expectancy_pred, everything())
```

Graphical Representation and Visualization

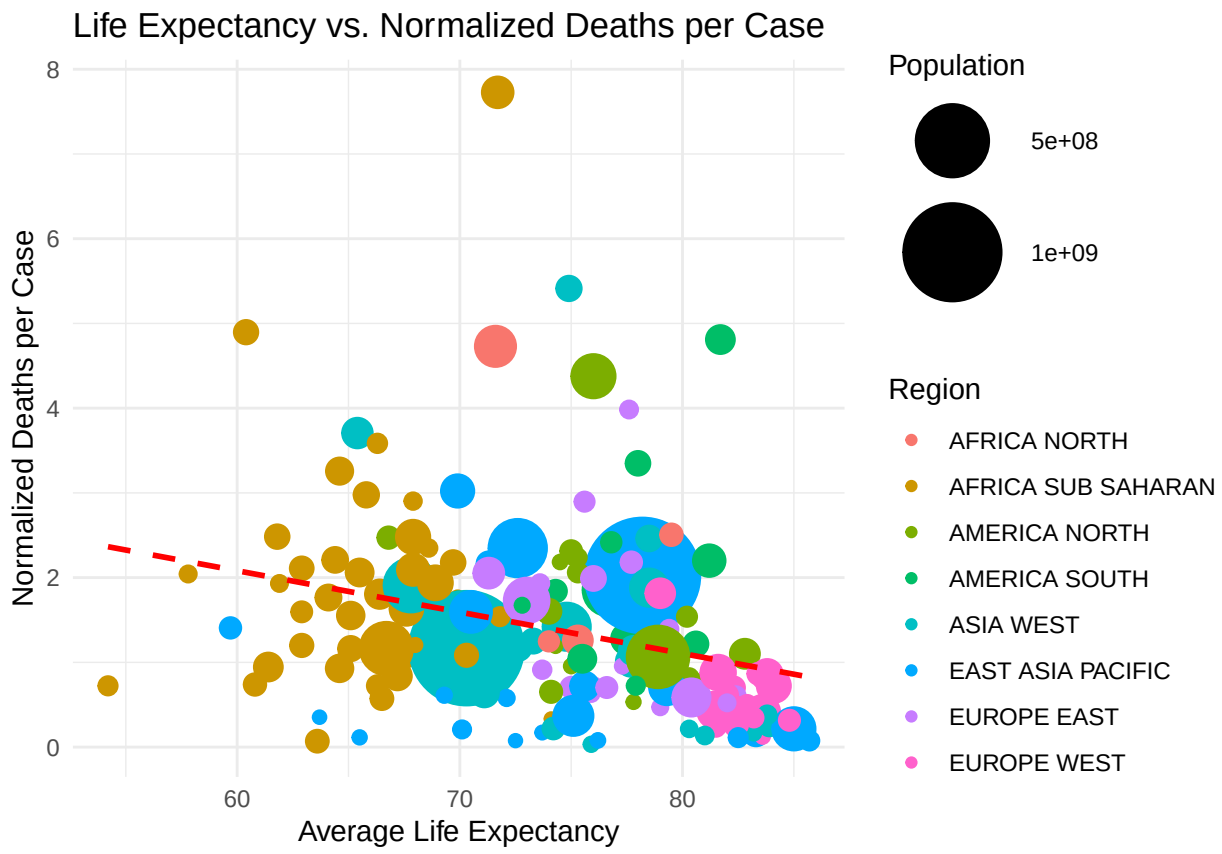
We are proceeding to visualize the relationships between income and normalized deaths per case, as well as life expectancy and normalized deaths per case. Each country's marker size in the plots is being scaled proportionally to its population, effectively conveying the relative impact of COVID-19. To enhance interpretability, countries are being color-coded based on their respective regions, aiding in the identification of trends and patterns. The visualizations feature dashed red lines that represent the projected relationships derived from the predictive models, enriching the plots with insights into expected patterns.

```
# Create a scatter plot for income vs. normalized deaths per case
income_vs_normalized_deaths =
  ggplot(combined_data_predictive, aes(x = Income,
                                       y = normalized_deaths_per_case)) +
  geom_point(aes(size = Population, color = Region)) +
  geom_line(aes(x = Income, y = Income_pred),
            linetype = "dashed", color = "red",
            linewidth = 1) +  # Add  red dashed line visualizing the predictive model
  scale_size_continuous(range = c(2, 20)) +
  labs(title = "Income vs. Normalized Deaths per Case",
       x = "Income Per Capita",
       y = "Normalized Deaths per Case") +
  theme_minimal()

# Create a scatter plot for life expectancy vs. normalized deaths per case
life_expectancy_vs_normalized_deaths =
  ggplot(combined_data_predictive, aes(x = Life_Expectancy,
                                       y = normalized_deaths_per_case)) +
  geom_point(aes(size = Population, color = Region)) +
  geom_line(aes(x = Life_Expectancy, y = Life_Expectancy_pred),
            linetype = "dashed", color = "red",
            linewidth = 1) +  # Add  red dashed line visualizing the predictive model
  scale_size_continuous(range = c(2, 20)) +
  labs(title = "Life Expectancy vs. Normalized Deaths per Case",
       x = "Average Life Expectancy",
       y = "Normalized Deaths per Case") +
```

```
  theme_minimal()

# Print the plots
print(income_vs_normalized_deaths)
```




```
print(life_expectancy_vs_normalized_deaths)
```



Conclusion: Unveiling Societal Realities

Our rigorous analysis, coupled with meticulous methodology and visualization prowess, paints a vivid picture of the intricate interplay between health outcomes and socio-economic variables.

While our study refrains from drawing sweeping conclusions, the data hint at an undeniable reality: the connection between resources and health care on COVID-19 mortality carries nuanced implications. The data suggest that areas burdened by austerity policies and socio-economic deprivation may face heightened vulnerability.

While cautious in our claims, it is hard to ignore the patterns that emerge. There's a stark indication that societal disparities, driven by economic inequalities, can manifest in dire health outcomes, especially in times of crisis.

Looking forward, our study ignites a call for comprehensive research to illuminate the underlying mechanisms of this connection. Such an understanding is crucial to inform decisions that can steer societies toward a more just and resilient future.

In the end, while our words remain cautious, our findings resonate with the undeniable truth that socio-economic inequalities can exacerbate health disparities. It's time to confront these realities head-on, with a commitment to creating a more inclusive and caring world for all.

Bias Assessment

In any analysis, a critical step is to acknowledge potential biases that could influence the results and interpretations. Here, we assess both explicit and implicit biases that may have impacted our study.

Explicit Exclusions

We want to be transparent about our data handling decisions. As part of our quality control measures, we removed certain countries from the analysis due to data irregularities. The countries explicitly removed are: **Andorra, Antarctica, Cabo Verde, Congo (Brazzaville), Congo (Kinshasa), Czechia, Diamond Princess, Dominica, Eswatini, Holy See, Korea, North, Korea, South, Kosovo, Kyrgyzstan, Laos, Liechtenstein, MS Zaandam, Marshall Islands, Micronesia, Monaco, Nauru, North Macedonia, Palau, Saint Kitts and Nevis, Saint Lucia, Saint Vincent and the Grenadines, San Marino, Slovakia, Summer Olympics 2020, Taiwan, Tuvalu, West Bank and Gaza, Winter Olympics 2022, Yemen**. Collectively, this accounts for a population of at least **35,290,600** individuals who were excluded from the analysis due to technical difficulties.

Sample Representation

It's important to note that our study relies on available data sources, which may not perfectly represent the entirety of the global population. Our conclusions are limited to the regions and countries covered in our dataset, which might not capture the full spectrum of socio-economic and health disparities.

Potential Survivorship Bias

One potential source of bias lies in the nature of the COVID-19 data itself. Our analysis is based on confirmed cases and deaths, but it's possible that repeated cases of COVID-19 in the same individual might have diminishing likelihood of resulting in death. This could lead to an underestimation of mortality rates, particularly in countries with high infection rates but lower death rates.

Socio-Economic Data Discrepancies

While we've endeavored to use the most reliable and up-to-date socio-economic indicators available, disparities in data collection and reporting across countries could introduce bias. Some regions might lack accurate reporting systems, potentially affecting the accuracy of the correlations we've explored.

A Call for Further Investigation

Despite our efforts to minimize biases, it's important to recognize that no analysis is devoid of potential limitations. To further enhance the accuracy and comprehensiveness of future analyses, we recommend a concerted effort to address these biases and explore correlations in diverse contexts.

In conclusion, our bias assessment underscores the need for ongoing research and meticulous data collection. By transparently acknowledging limitations, we pave the way for more comprehensive investigations that can ultimately contribute to a deeper understanding of the relationship between socio-economic factors and COVID-19 outcomes.