Personal Project Quarto

Daniel Liu, Jason Ficorilli

Global Health - Where's My Water?

As a part of a recent project launched by MoreLife Ltd., we were recruited to conduct some preliminary data analysis regarding water access percent versus life expectancy, as well as a closer look at life expectancy results in general, to see if there are any meaningful relationships we can draw.

Specifically, we'll first be working towards seeing if there's a significant relationship between water access percent and life expectancy when the countries in the dataset are stratified into five regions, and analyzing the resulting visualizations. Depending on the results, MoreLife Ltd. may decide to allocate more resources towards providing safe water access.

Then, we'll aim to focus on life expectancy in small subsets of countries, to determine which characteristics most drastically separate countries with high and low life expectancies. Due to MoreLife's extensive network with other federal agencies, understanding the effects of factors such as GDP on life expectancy can be valuable information for them, as well.

[Describe what attributes you'll focus your analysis on (mention if they are part of your data sets or if you created them out of your data sets).]

Dataset

Provenance

The primary dataset that we will be using to complete this project will have come from Kaggle, titled "Global Health and Development (2012-2021)". It was created by Martina Galasso, who collected over 60 datasets from the World Health Organization and the World Bank to create it.[1]

According to Galasso, the purpose of the dataset was to provide information for "researchers, analysts, and policymakers" who were interested in exploring the relationship between the factors present in the dataset.

[1] Galasso (2024)

FAIR and CARE

Findability can be ensured due to the data source, Kaggle, being completely publicly available. Kaggle is already a very common tool for dataset search, and simply filtering by name can bring up our dataset.

Accessibility is also confirmed for a similar reason, where anybody who finds the dataset on Kaggle should be able to directly download it, and any other related files.

Interoperability can be ensured because the dataset currently uses very easily-readable column names, as well as a pretty simple format that is hard to misinterpret. If there are any concerns, they can also be resolved during data wrangling.

Reusability is confirmed through the Kaggle website - if the dataset is found, all necessary metadata is located towards the bottom of the page.

Collective benefit can be seen by the fact that a very diverse group of countries is represented in the dataset. It is difficult to engage in any biases considering the fact that all of the metrics included in the dataset are purely objective.

Authority to Control is debatable. The Indigenous people whose data was used in this dataset was gathered by the World Health Organization and the World Bank. All of this information is completely publicly available.

Responsibility and **Ethics** can also be ensured, for a similar reason that collective benefit can. There's little to no room for bias in this dataset, considering all of the metrics are purely objective. Historical context is not necessarily reflected, but also would not have any tangible effect on the data itself (and the same goes with cultural values). The interpretation of this data can certainly be affected by historical context, which we will definitely consider as we move forward.

Data cleaning

To better understand the structure of our main dataset, let's first take a look at the head, and a few example columns (the rest of the 29 indicators are visible in the original dataset, but for the sake of aesthetics, only the first few will be displayed).

Table 1: Raw View of Global Health Data

Country	Country_Code	Year	Fertility_Rate	Urban_Population_Percent
Afghanistan	AFG	2012	5.830	24.160
Afghanistan	AFG	2013	5.696	24.373
Afghanistan	AFG	2014	5.560	24.587
Afghanistan	AFG	2015	5.405	24.803
Afghanistan	AFG	2016	5.262	25.020
Afghanistan	AFG	2017	5.129	25.250

As we can see in Table 1, the data seems to be decently tidy - at first glance, the columns are well-named, and the numbers don't include any commas that would make computation difficult. The Country_Code column also is helpful for code brevity.

For this part of the research paper, we are only concerned with Country, Country_Code, Year, Water_Access_Percent, and Life_Expectancy. As such, we can get rid of all of the other columns for now and conduct our operations on a new dataframe, shown below, in Table 2.

Table 2: Relevant Columns of Global Health Data

Country	Country_Code	Year	Water_Access_Percent	Life_Expectancy
Afghanistan	AFG	2012	21.12400	61.923
Afghanistan	AFG	2013	22.03447	62.417
Afghanistan	AFG	2014	22.94430	62.545
Afghanistan	AFG	2015	23.85359	62.659
Afghanistan	AFG	2016	24.76222	63.136
Afghanistan	AFG	2017	25.67142	63.016

- regardless, we can move forward with some actual discovery.

Part One: Evaluating the Relationship Between Water Access Percent and Life Expectancy Across Region

Preliminary Investigation

Firstly, it might be beneficial to take a look and see if there's a relationship at all between life expectancy and water access percent, across all of the countries.

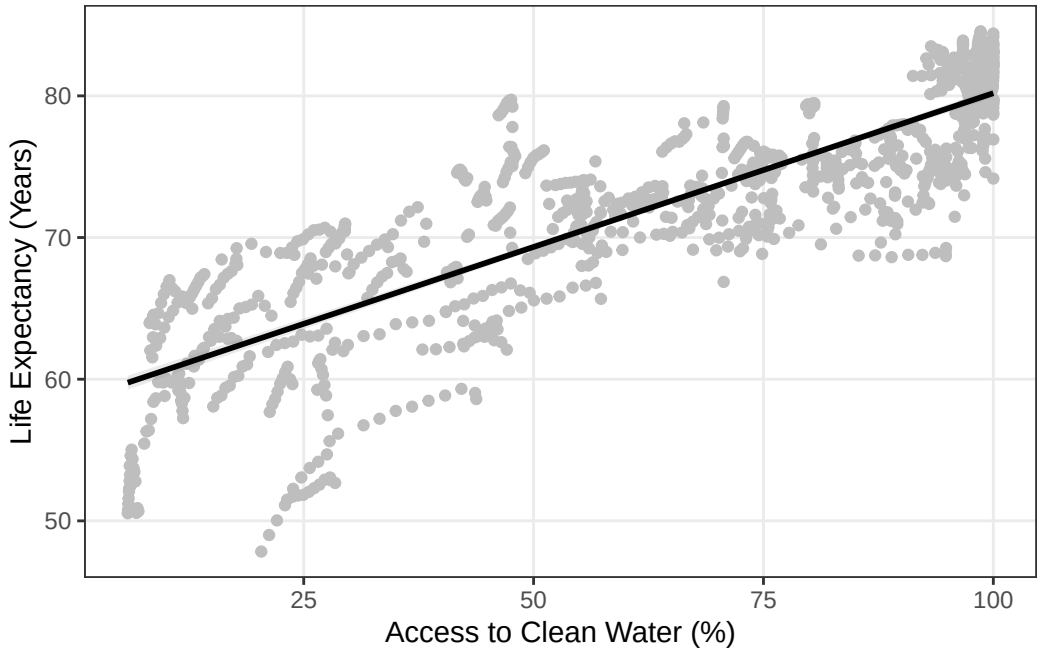


Figure 1: Scatterplot of Water Access Percent Between 2012-2021

According to Figure 1, apparently, there does seem to be a decently strong correlation between access to clean water and life expectancy. One thing to note is that there seem to be a lot of countries closely clustered toward 100% access to clean water, and most of the associated life expectancies seem to be clustered around 82% - 85%.

For countries with lower water access percentages, the distribution of associated life expectancies is a lot more spread-out, and thus, would probably be harder to predict. Additionally, if we look closely, we can see what can seem to be little clusters of dots that, presumably, are all from the same country, but over the years (these are more visible toward the left of the graph).

As this graph stands, though, it isn't very helpful in trying to identify why any of these differences in access to clean water or life expectancy may exist. More specifically, MoreLife Ltd. specializes in on-the-ground work, so let's take a look at how region may impact this graph.

Does Region Affect Water Access Percent vs. Life Expectancy?

Data Wrangling

First, let's take a look at our secondary dataset, which is a collection of M49 codes from the United Nations Statistical Division [2]. It contains standardized country codes, with region

[2]United Nations Secretariat (1999)

information, as well as ISO-alpha3 and ISO-alpha2 codes (Table 3).

Table 3: Raw View of UNSD Code Data

Global Code	Global Name	Region Code	Region Name	Sub-region Code
1	World	2	Africa	15
1	World	2	Africa	15
1	World	2	Africa	15
1	World	2	Africa	15
1	World	2	Africa	15
1	World	2	Africa	15

For the purposes of our data, we'll only be needing the following columns: `Region Name`, and `ISO-alpha3 Code`. Let's filter by just those two columns (visible in Table 4, below).

Table 4: Relevant Columns of UNSD Code Data

Region Name	ISO-alpha3 Code
Africa	DZA
Africa	EGY
Africa	LBY
Africa	MAR
Africa	SDN
Africa	TUN

Now, we can begin with merging the two datasets via the `ISO-alpha3 Code` and `Country Code` columns, which both mean the same thing.

Table 5: Raw View of Merged Global Health and UNSD Code Data

Country	Country_Code	Year	Water_Access_Percent	Life_Expectancy	Region Name
Afghanistan	AFG	2012	21.12400	61.923	Asia
Afghanistan	AFG	2013	22.03447	62.417	Asia
Afghanistan	AFG	2014	22.94430	62.545	Asia
Afghanistan	AFG	2015	23.85359	62.659	Asia
Afghanistan	AFG	2016	24.76222	63.136	Asia
Afghanistan	AFG	2017	25.67142	63.016	Asia

Now, we have our original dataset with all of our necessary information, as well as an extra column that includes each specific country's Region (Table 5). We can use this extra column to help see if its eventually stratification, at all, affects our visualization, or at least, gives us some insight as to what specific Regions MoreLife Ltd. should focus on in their research. Let's begin.

Summary Statistics

First, let's see if there's anything interesting when it comes to summary statistics for individual regions. We saw summary statistics for the whole dataset, but actually splitting it up by some metric is much more useful. Let's see how that looks.

Table 6: Summary Statistics of Merged Global Health and UNSD Code Data

Region Name	Mean	Median	SD	Min	Max	Range
Africa	28.56	23.68	20.74	5.86	76.97	71.11
Americas	69.07	64.15	18.86	41.68	99.04	57.36
Asia	67.32	68.42	26.72	14.59	100.00	85.41
Europe	93.62	96.93	8.30	66.38	100.00	33.62
Oceania	47.64	41.16	31.67	8.24	100.00	91.76

At first glance, when looking at Figure 7, there is a very large difference among the five regions in almost every metric shown above, so we will discuss each one. First, the **Mean** `Water_Access_Percent` is highest in Europe, and lowest in Africa. They differ by nearly 65 percentage points. America and Asia both have relatively similar mean `Water_Access_Percent`, and Oceania falls slightly lower. The **Median** statistic reflects these observations.

However, **Standard Deviation** provides some interesting points. Firstly, Europe seems to have the lowest standard deviation, which actually matches what we saw earlier with the graph (where the higher water access percents all seemed to be clustered around a similar life expectancy, while lower water access percents tended to have greater variation in terms of life expectancy). Additionally, even though Oceania wasn't particularly remarkable in terms of its mean `Water_Access_Percent`, it has the highest standard deviation. In fact, its standard deviation is nearly 2/3 of its mean, which clearly indicates a huge range of `Water_Access_Percent` values.

Oceania's relatively large standard deviation is also reflected in its **Min** and **Max**, which, by far, has the largest difference (which we can also see through its high **Range**).

Judging from Figure 7, we can suggest to MoreLife Ltd. to focus on Africa, given that they have the lowest average water access percentage, and to potentially pay closer attention to regions such as Oceania, since this range indicates some areas being extremely "wealthy", and other areas being more neglected.

However, to get an even better understanding of how this information actually affects the *relationship* between water access percentage and life expectancy, let's create a graph like before, but now reflecting region.

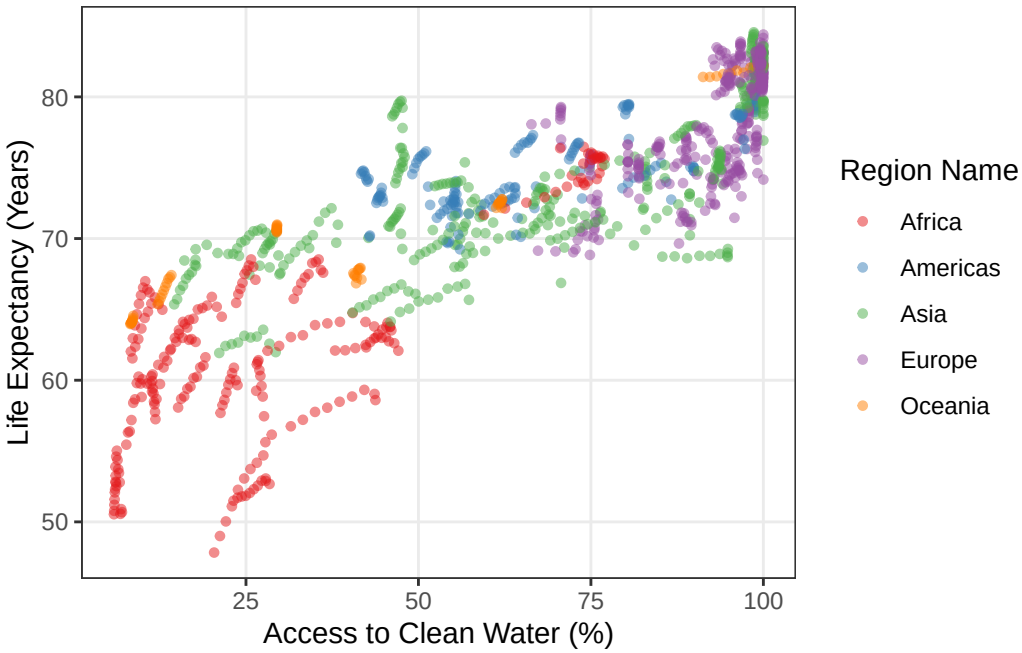


Figure 2: Scatterplot of Water Access Percent vs. Life Expectancy Between 2012-2021, Colored By Region

What we can see in Figure 2 is actually quite a bit of interesting information regarding the distribution of each region's water access percent. Toward the right and top side of the graph, we can now see that the cluster of points with nearly 100% `Water_Access_Percent` are composed mainly of some European and Asian countries, with a few in Oceania. Interestingly enough, even though America had the second-highest mean `Water_Access_Percent`, they're not as clustered toward the top as some Asian countries, and are more spread out.

Additionally, we can see what we noted previously, where Oceania was much more spread out than any of the other groups. Although there aren't many, there are little clusters of orange dots that range anywhere from about 10% water access to nearly 100% water access (this also reflects what we saw with min/max summary statistics).

Other regions such as Africa and the Americas are decently clustered around the same area. However, upon further review, it is a little bit difficult to tell where the points are clustered in a scatterplot context. This is especially true considering Oceania's distribution being very modal, and makes it seem less impactful than it is. In order to solve this, we can create a graph of contours around points to better understand the distribution of regions.

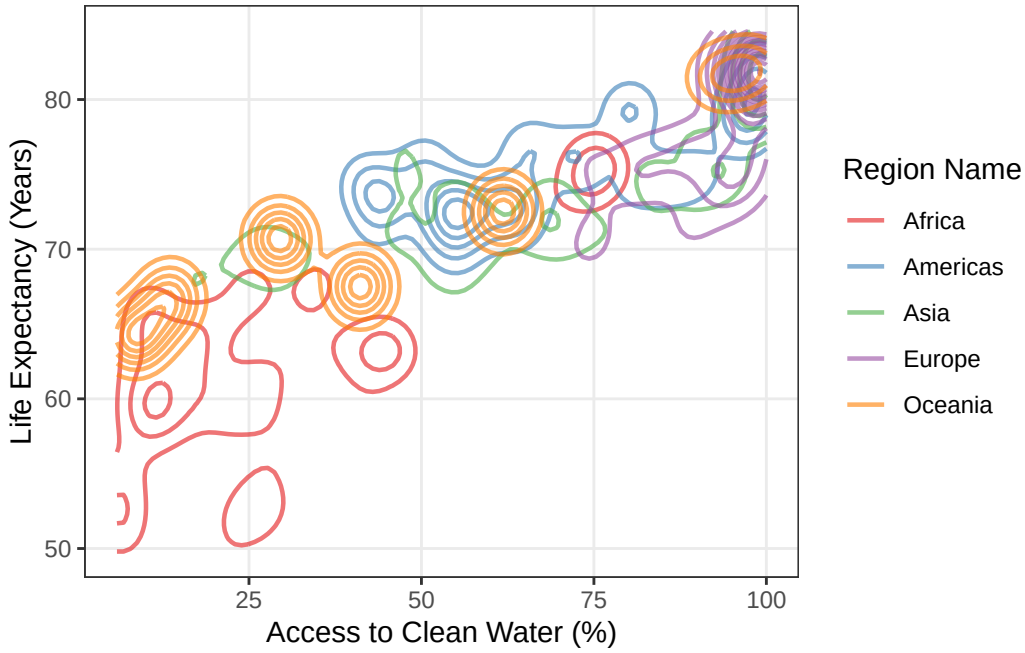


Figure 3: Contour Plot of Water Access Percent vs. Life Expectancy Between 2012-2021, Colored By Region

Finally, we have a graph (Figure 3) that best represents the distribution of region data. Specifically, we can now see the Oceania data points much more clearly, and see that they're specifically clustered around 10%, 26%, 40%, 60%, and 90% water access percentage, approximately. We can also see a huge cluster of African countries around 75% water access percentage that was decently hard to spot in the scatterplot.

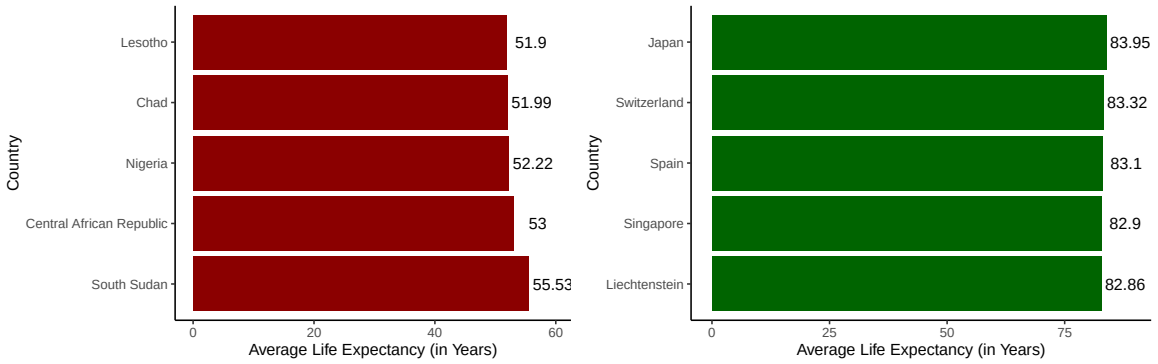
Overall, we can now see that the correlation between access to clean water percentage is very strong in general, and although the influence of regions on this correlation is also decently strong, it has a lot more discrepancies than one might assume. However, MoreLife Ltd. is very much interested in further exploring life expectancy as an attribute, as well as region, since they seem to show a lot of promise for future discovery.

Part Two: Evaluating the Temporal Characteristics of Countries with High and Low Life Expectancies

After considering the relationship between life expectancy and safe water access across all available countries within the data set, it is reasonable to question what attributes separate countries with high and low life expectancy. With this in mind, our focus can narrow to isolating small subsets of countries with vastly differing life expectancy rates and comparing different characteristics about their treatment of healthcare over time.

Highest and Lowest Life Expectancy Countries

In order to identify countries to investigate further, we demonstrate the top five countries with the highest and lowest average life expectancy measures from 2012-2021, as shown in each respective bar chart.



(a) Top Five Highest Average Life Expectancy Countries

(b) Top Five Lowest Average Life Expectancy Countries

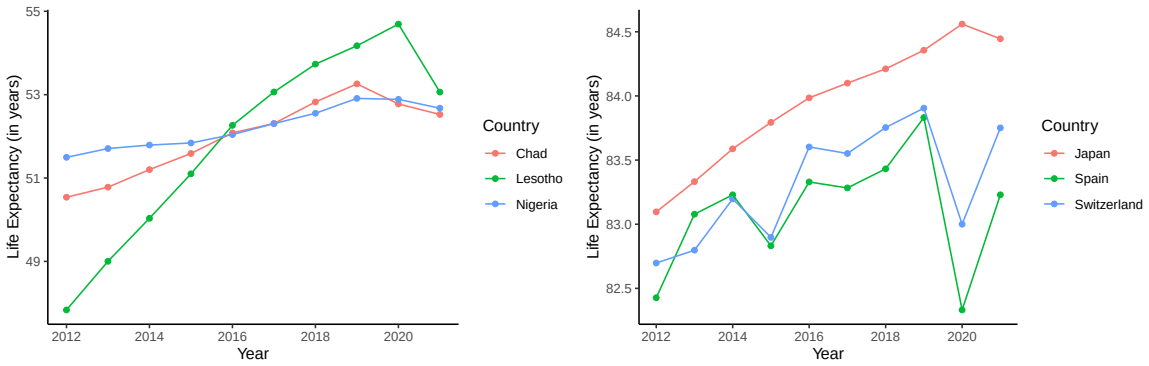
Figure 4: Top Five Countries with the Highest and Lowest Average Life Expectancy Measures from 2012-2021

As captured in Figure 4, the differences in the highest and lowest life expectancy measures is stark. Over the period of time from 2012-2021, a person would have been expected to live an average of 32 years longer as a Japanese resident compared to having lived in Lesotho. Geographical differences are also apparent: the top five countries with the lowest average life expectancy are in Africa, while the top five highest average life expectancy are in Asia and Europe. To further explore the differences between these types of countries, the three highest (**Japan**, **Switzerland**, and **Spain**) and three lowest (**Lesotho**, **Chad**, and **Nigeria**) average life expectancy countries will be primarily evaluated.

Considering such differences in average life expectancy from 2012-2021, the year-by-year basis of each country's life expectancy can be explored further, as shown in Figure 5.

The general life expectancy changes for each country from 2012-2021 reveal both similarities and differences between countries with high and low measures during that time. The three countries with lower life expectancy (Chad, Lesotho, Nigeria) all exhibited similar upward growth from 2012-2020, with Lesotho eclipsing the other countries in 2016 and adding over 5 years to its overall life expectancy. Although all three African countries showed consistent growth, Spain, Switzerland, and Japan were not as similar. While Japan held a consistent increase in life expectancy, both Spain and Switzerland saw decreases from 2014-2015 and, most notably from 2019-2020.

Although each country's trends were generally upward over time, a notable similarity was a decrease in life expectancy across every country within a part of the time period 2019-2021. This is likely the result of the COVID-19 global pandemic, which contributed to an estimated

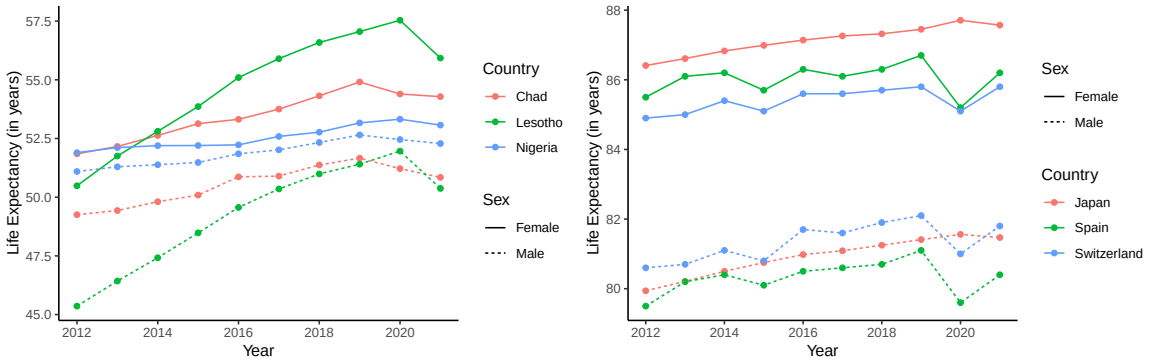


(a) Change in Life Expectancy for Chad, Lesotho, and Nigeria

(b) Change in Life Expectancy for Japan, Spain, and Switzerland

Figure 5: Life Expectancy Measures for Select Countries from 2012-2021

18.2 million global deaths from 2020-2021.[3] Interestingly, while Japan, Chad, Nigeria, and Lesotho all saw decreases in life expectancy from 2020-2021, Spain and Switzerland had considerable increases. Further, among the countries that did exhibit a loss during this time, Lesotho decreased the largest from 2020-2021, losing 1.63 years, while Spain lost 1.1 years from 2019-2020. The similarities and differences in life expectancy among these countries are further exemplified when compared on the basis of sex.



(a) Change in Life Expectancy for Chad, Lesotho, and Nigeria

(b) Change in Life Expectancy for Japan, Spain, and Switzerland

Figure 6: Life Expectancy Measures for Select Countries from 2012-2021 by Sex

The sex-based visuals in Figure 6 display differing characteristics from the general life expectancy trends. Firstly, while most countries have had a higher female life expectancy of at least three years compared to the male rate throughout this time, Nigeria's life expectancy was considerably close for both men and women, with nearly identical values for several years. Further, of the European and Asian countries, women were expected to live nearly four years longer than men on average. These trends show that, while the magnitude of life

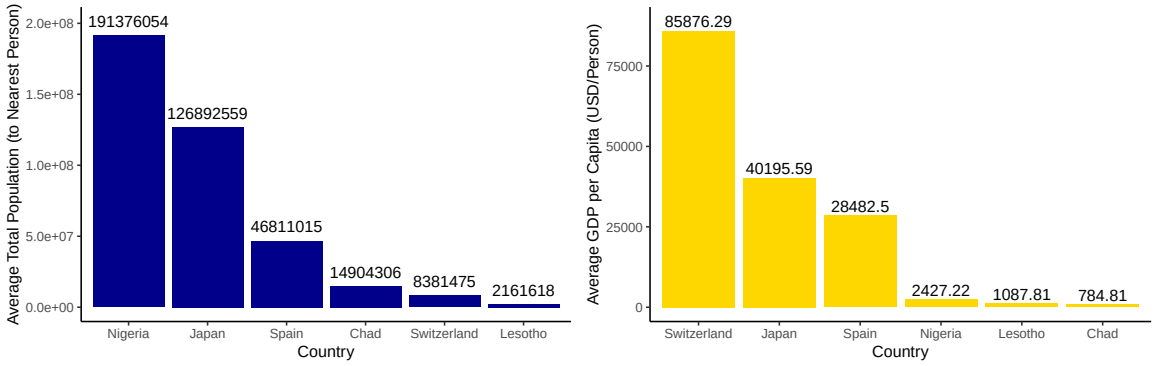
[3]Wang et al. (2022)

expectancy for the three African countries and three European/Asian countries were vastly different, these regions all exhibited some similarities in the recent developments of their life expectancy.

After evaluating the temporal elements of the six countries' life expectancy values, the possible attributing factors to these changes can be addressed. While many factors influence the life expectancy measure for a country, several of which are beyond the scope of this project, the global health data set provides additional characteristics recorded over the period from 2012-2021 that further illustrate the differences between these countries.

Differences in Population and Healthcare Expenditure

When looking at why a country might have experienced changes in the expected lifetime of their residents, it is important to consider the demographic and economic factors enabling a country to improve its population's health. Although two countries might have similar life expectancy values, vastly different government support systems and national finances could underlie their capabilities to perform effective public health initiatives. These comparisons provide necessary insight for evaluating how multiple countries have observed fluctuation in resident health over time. Firstly, we will discuss the differences of population and GDP between the six countries considered thus far.



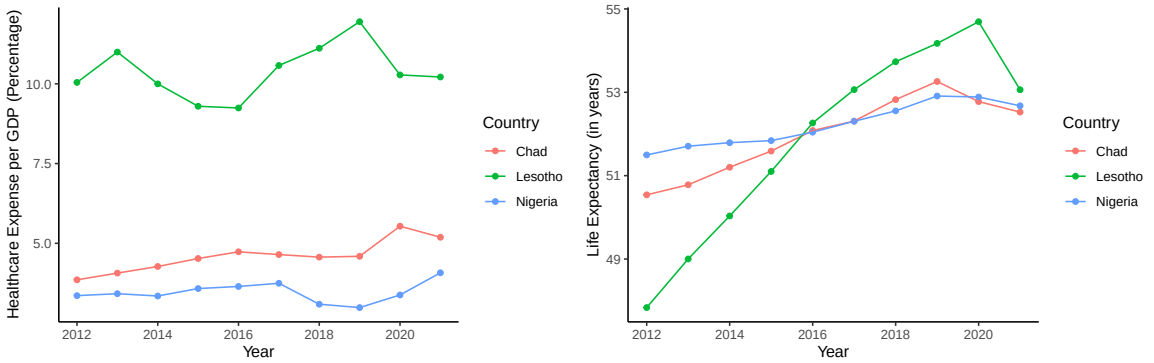
(a) Average Total Population

(b) Average GDP per Capita

Figure 7: Average Total Population and Gross Domestic Product per Capita for Chad, Japan, Lesotho, Nigeria, Spain, and Switzerland from 2012-2021

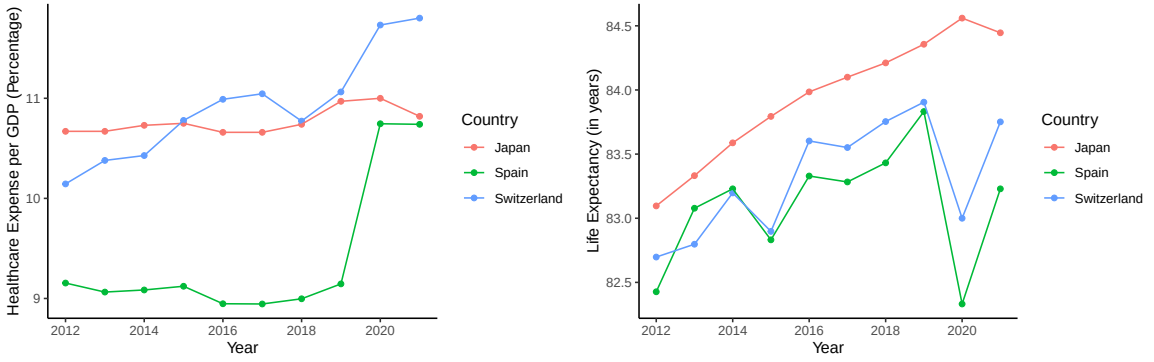
The visuals of GDP and population size in Figure 7 set the context for considering the life expectancy measures for each country. As shown in the average total population bar chart, Nigeria had a considerably larger population than all the other countries, with Switzerland having the second lowest followed by Lesotho. However, Switzerland had more than double the average GDP per capita than Japan and over 19 times more than the lowest three life expectancy countries combined. Further, Spain had over six times the total average GDP per capita of the three African countries, with Chad having only $784.81 per person on average from 2012-2021.

Although national wealth is not the sole factor determining how long a country's residents are expected to live, it can represent a country's ability to provide adequate expenditure and government support for the healthcare needs of its population. In particular, it can reflect the financial support available for a country to combat serious diseases or mitigate air pollution, both of which could significantly decrease a country's life expectancy if widespread. This leads us to consider the comparison of each country's healthcare expenditure over time.



(a) Healthcare Expenditures per GDP (b) Change in Life Expectancy

Figure 8: Healthcare Expenditures per Gross Domestic Product and Life Expectancy from 2012-2021 for Chad, Lesotho, and Nigeria



(a) Healthcare Expenditures per GDP (b) Change in Life Expectancy

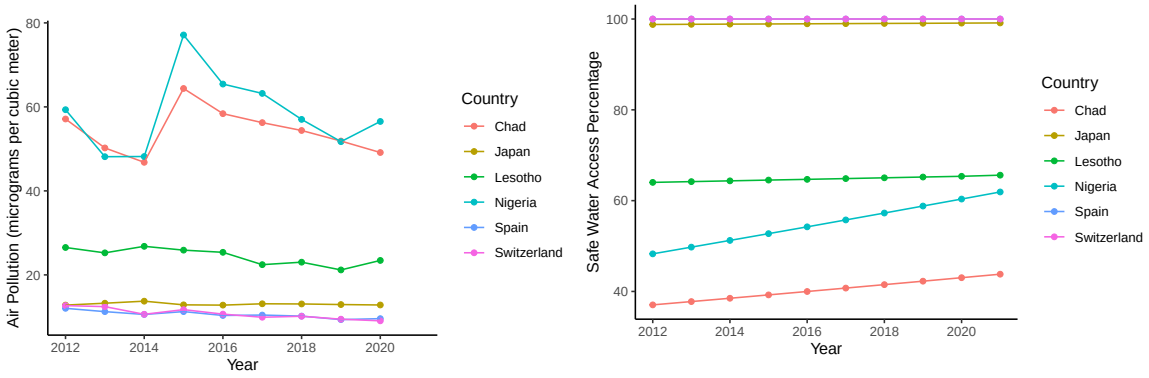
Figure 9: Healthcare Expenditures per Gross Domestic Product and Life Expectancy from 2012-2021 for Japan, Spain, and Switzerland

The direct comparison of the countries' healthcare expenditure to life expectancy changes reveals conflicting characteristics about the efficacy of large national healthcare spending. In Figure 8, when observing the three African countries, Lesotho's large expenditure percentage of its GDP towards healthcare seems to be a worthwhile investment as its life expectancy has risen considerably over the years relative to similar average expectancy countries. However, in Figure 9, when observing the European and Asian countries, Spain's very low expenditure relative to Switzerland and Japan while maintaining similar life expectancy may suggest the opposite.

Either way, a notable observation is the magnitude of the expenditure percentages of GDP across the countries. Yearly, Chad and Nigeria combined spent less than Lesotho on healthcare: roughly only 3-4% each. This is less than double the rate of Lesotho, Japan, and Switzerland respectively, which may reflect the countries' slowly increasing life expectancy relative to Lesotho.

Differences in Air Pollution and Safe Water Access

Although a country can spend significant finances relative to its GDP to improve its population's healthcare treatment, this expenditure must also combat sources that put its population at a health risk. To consider these possible factors, we observe patterns of air pollution exposure and safe water access relative to life expectancy for the six countries.



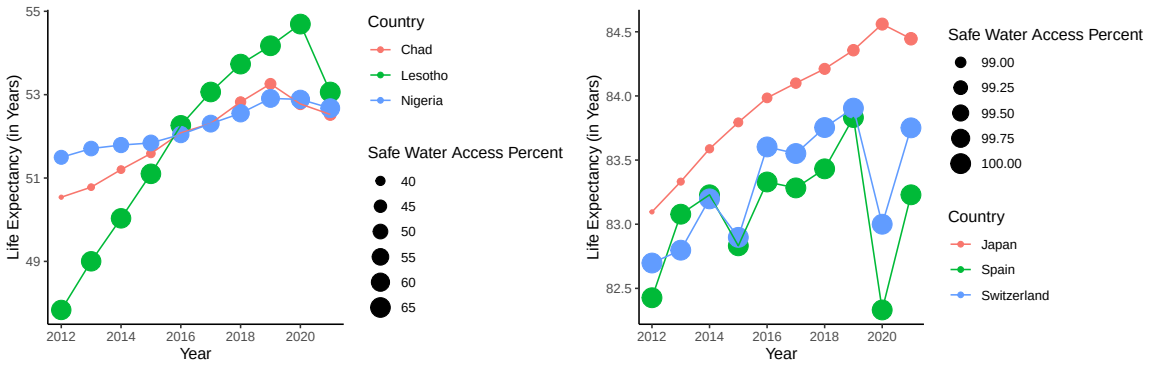
(a) Change in Air Pollution Levels from 2012-2020 (b) Change in Safe Water Access Percentage from 2012-2021

Figure 10: Air Pollution Levels and Safe Water Access Percentage from 2012-2021 for Chad, Japan, Lesotho, Nigeria, Spain, and Switzerland

Figure 10 and 11's pollution and safe water access visuals reveal considerable insights into potential causes for the life expectancy discrepancies between the six countries. In the graphs, the three African countries have air pollution levels considerably higher than the European and Asian countries. Similarly, over a third of the population for each African country does not have safe access to water, while nearly 100% of the other countries do. These results highlight the major disparities in population quality of life between countries with high and low life expectancy. Although there are promising developments with increasing rates of safe water access and decreasing air pollution levels for Chad and Nigeria, the high magnitudes relative to a country like Lesotho likely remain influential factors in harming overall population life expectation.

Conclusion

Through this analysis, we have evaluated the different dimensions of global life expectancy and its relationship with safe water access. In part one, after preparing our data, we eval-



(a) Chad, Lesotho, and Nigeria

(b) Japan, Spain, and Switzerland

Figure 11: Life Expectancy and Safe Water Access Percentage from 2012-2021 for Chad, Japan, Lesotho, Nigeria, Spain, and Switzerland

uated a possible relationship between a country's safe water access and its population life expectancy. Here, we found that, on average, life expectancy for a country tends to increase given an increase in safe water access percentage. In part two, we identified countries with high and low average life expectancy. Through analyzing the relationships between life expectancy and gender, healthcare expenditure, and air pollution levels, we identified the similarities and differences in such factors attributing to population health. In particular, we found that the subset of countries with the lowest average life expectancy (**Chad**, **Lesotho**, and **Nigeria**) had considerably higher levels of air pollution and less population access to safe water, which could be significantly influencing their countries' vastly lower life expectancy relative to the highest average countries (**Japan**, **Span**, and **Switzerland**).

Considering these results, it is important to consider if countries' healthcare expenditure truly impacts the improvement of overall health outcomes. For Chad and Nigeria in particular, their high air pollution levels and low safe water access amid small healthcare expenditure per GDP (relative to high average life expectancy countries) signals an important role. Further, Lesotho saw a considerable increase in its life expectancy over its period of spending roughly 10% of its GDP on healthcare. However, a causal relationship cannot be determined from this analysis alone and must be considered in the context of many more countries' observations.

Code Appendix

```
# Style Guide: BOAST

library(ggplot2)
library(tidyverse)
library(readr)
library(dplyr)
```

```r
library(kableExtra)

#Load main data set
healthData <- read.csv(
  url(
    "https://raw.githubusercontent.com/Stat184-Fall2024/Sec1_FP_DanielLiu_JasonFicorilli/ref
    )
  )

#Load global regions data set
urlFileRegion = "https://raw.githubusercontent.com/Stat184-Fall2024/Sec1_FP_DanielLiu_JasonF

dataRegion <- read_csv("../data/UNSD_codes.csv") # Getting the file from the working directo

#Select all columns from 2012-2021 table except year and country code
countryAverageDF <- healthData %>%
  select(
    !c(Year, Country_Code)
    ) #Remove year and country code columns

#Create data frame of averaged columns
countryAverageDF <- aggregate( #aggregate columns to be averages from 2012-2021
  x = countryAverageDF[, 2:27],
  by = list(countryAverageDF$Country),
  FUN = mean
) %>%
  rename("Country" = "Group.1")

#Create function to display tables with kable
displayTable <- function(table) {
  newTable <- table %>%
    head() %>% # Displaying just the head for aesthetics
    knitr::kable() %>% # Styling - using kable striped
    kable_styling(
      latex_options = c("striped", "hold_position") # Specifically choosing striped for aest
    )
  return(newTable) # Return call
}
# View the raw data of our Global Health dataset
healthDataRaw <- healthData %>%
  select(Country, Country_Code, Year, Fertility_Rate, Urban_Population_Percent) # Selecting

healthDataRaw %>% displayTable()
# View only the relevant columns from our Global health dataset
```

```r
dataWater <- healthData %>% select(Country, Country_Code, Year, Water_Access_Percent, Life_E
dataWater %>% displayTable()
# Scatterplot with trend line of water access percentage vs. life expectancy
ggplot(
  data = dataWater,
  aes(
    x = Water_Access_Percent,
    y = Life_Expectancy)) +
  geom_point(color = "grey") + # Grey points, to help with making them less visible
  geom_smooth(method = "lm", # Adding trend line
              color = "black", # Making the trend line black to make it more clear
              se = TRUE,
              alpha = 0.2) + # Making the dots a little bit smaller for visibility
  theme_bw() + # Black and white theming
  labs(
    x = "Access to Clean Water (%)",
    y = "Life Expectancy (Years)",
  ) +
  theme(
    panel.grid.minor = element_blank() # Making the minor grid lines less visible since thei
  )
# Viewing the raw data for our UNSD Code Data
dataRegionRaw <- dataRegion %>%
  select("Global Code", "Global Name", "Region Code", "Region Name", "Sub-region Code") # Se

dataRegionRaw %>% displayTable()
# Table of relevant columns from USD Code Data
dataRegionRelevant <- dataRegion %>%
  select("Region Name", "ISO-alpha3 Code") # Selecting only the relevant columns from our da

dataRegionRelevant %>% displayTable()
# Table of Global Health merged with UNSD Code Data
dataRegionMerged <- left_join( # Performing a left join to prioritize our Global Health data
  x = dataWater,
  y = dataRegionRelevant,
  by = join_by("Country_Code" == "ISO-alpha3 Code") # Join condition
)

dataRegionMerged %>% displayTable()
# Table of summary statistics for Global Health UNSD Code merged table
dataRegionSummary <- dataRegionMerged %>%
  group_by(
    `Region Name`
  ) %>% # Specifically grouping my region name to compare the different regions
```

```r
  summarize(
    Mean = round(mean(Water_Access_Percent, na.rm = TRUE), 2), # Mean, median, SD, Min, and
    Median = round(median(Water_Access_Percent, na.rm = TRUE), 2),
    SD = round(sd(Water_Access_Percent, na.rm = TRUE), 2),
    Min = round(min(Water_Access_Percent, na.rm = TRUE), 2),
    Max = round(max(Water_Access_Percent, na.rm = TRUE), 2)
  ) %>%
  mutate(
    Range = round(Max - Min, 2)
  ) # Rounding all of the values to 2 decimal points for readability

dataRegionSummary %>% displayTable()
# Scatterplot of water access percent vs. life expectancy with region considered
ggplot(
  data = dataRegionMerged, # Data
  aes(
    x = Water_Access_Percent,
    y = Life_Expectancy,
    color = `Region Name`)) +
  geom_point(
    alpha = 0.5,
    size = 1.2) + # Making the dots less opaque, but still large enough discerned
  scale_color_brewer(palette = "Set1") + # Selecting a nice color palette that can distingui
  theme_bw() +
  labs(
    x = "Access to Clean Water (%)",
    y = "Life Expectancy (Years)",
    color = "Region Name", # Actually coloring the dots by region
    fill = "Region Name" # ^
  ) +
  theme(
    panel.grid.minor = element_blank()
  )
# Contour plot of the merged data comparing water access percentage and life expectancy
ggplot(
  data = dataRegionMerged,
  aes(
    x = Water_Access_Percent,
    y = Life_Expectancy,
    color = `Region Name`)) +
  geom_density_2d(aes(color = `Region Name`), # Actually using a contour plot to make distri
                  linewidth = 0.8, # This and below - aesthetics to make the contours better
                  alpha = 0.6,
                  bins = 15,
```

```
                  h = c(12, 6)) +
  scale_color_brewer(palette = "Set1") + # Choosing a nice color palette to actually disting
  theme_bw() +
  labs(
    x = "Access to Clean Water (%)",
    y = "Life Expectancy (Years)",
    color = "Region Name", # Actually coloring by region
    fill = "Region Name"
  ) +
  theme(
    panel.grid.minor = element_blank()
  )
#Horizontal bar chart for the five lowest average life expectancy countries
ggplot(
  data = countryAverageDF %>% arrange(Life_Expectancy) %>% head(5), #get top 5 countries
  mapping = aes(
    x = Life_Expectancy,
    y = reorder(Country, -Life_Expectancy)) #reorder bars to be descending
) +
  geom_col(fill="darkred")+
  geom_text(
    aes(label = signif(Life_Expectancy, 4)),
    nudge_x = 4) +
  #add respective values (4 sig figs) next to bars
  labs(
    x = "Average Life Expectancy (in Years)",
    y = "Country"
  ) +
  theme_classic()


#Horizontal bar chart for the five highest average life expectancy countries
ggplot(
  data = countryAverageDF %>% arrange(desc(Life_Expectancy)) %>% head(5), #get top 5
  mapping = aes(
    x = Life_Expectancy,
    y = reorder(Country, Life_Expectancy)) #reorder bars to be descending
) +
  geom_col(fill="darkgreen")+
  geom_text(aes(label = signif(Life_Expectancy, 4)), nudge_x = 5) +
  #add respective values (4 sig figs) next to bars
  labs(
    x = "Average Life Expectancy (in Years)",
    y = "Country"
```

```
  ) +
  theme_classic()
#Life expectancy line chart for low 3 countries over time
ggplot(
  data = healthData %>% filter(Country=="Nigeria" | Country=="Lesotho" | Country =="Chad"),
  mapping = aes( #select dimensions to plot
    x = Year,
    y = Life_Expectancy,
    color = Country
  )
) +
  geom_point()+
  geom_line()+
  labs( #select labels for plot
    x = "Year",
    y = "Life Expectancy (in years)",
    color = "Country"
  ) +
  theme_classic() #change theme (remove grid lines)


#Life expectancy line chart for high 3 countries over time
ggplot(
  data = healthData %>% filter(Country=="Japan" | Country=="Spain" | Country =="Switzerland"
  mapping = aes( #select dimensions to plot
    x = Year,
    y = Life_Expectancy,
    color = Country
  )
) +
  geom_point()+
  geom_line()+
  labs( #select labels for plot
    x = "Year",
    y = "Life Expectancy (in years)",
    color = "Country",
  ) +
  theme_classic() #change theme (remove grid lines)
# Life expectancy line chart for low 3 countries, split on sex
ggplot(
  data = healthData %>%
    filter(Country=="Nigeria" | Country=="Lesotho" | Country =="Chad") %>%
    select(Country, Year, Life_Expectancy_Male, Life_Expectancy_Female) %>%
    rename("Male" = "Life_Expectancy_Male") %>%
```

```
    rename("Female" = "Life_Expectancy_Female") %>%
    pivot_longer(
      cols = c("Male", "Female"),
      names_to = "Sex",
      values_to = "Life_Expectancy"
    ),
  mapping = aes( #select dimensions to plot
    x = Year,
    y = Life_Expectancy,
    color = Country
  )
) +
  geom_point()+
  geom_line(aes( #creating different line types based 'Sex'
    linetype = Sex
  ))+
  labs( #select labels for plot
    x = "Year",
    y = "Life Expectancy (in years)",
    color = "Country"
  ) +
  theme_classic() #change theme (remove grid lines)


#Life expectancy line chart for high 3 countries over time
ggplot(
  data = healthData %>%
    filter(Country=="Japan" | Country=="Spain" | Country =="Switzerland") %>%
    select(Country, Year, Life_Expectancy_Male, Life_Expectancy_Female) %>%
    rename("Male" = "Life_Expectancy_Male") %>%
    rename("Female" = "Life_Expectancy_Female") %>%
    pivot_longer(
      cols = c("Male", "Female"),
      names_to = "Sex",
      values_to = "Life_Expectancy"
    ),
  mapping = aes( #select dimensions to plot
    x = Year,
    y = Life_Expectancy,
    color = Country
  )
) +
  geom_point()+
  geom_line(aes( #creating different line types based 'Sex'
```

```
    linetype = Sex
  ))+
  labs( #select labels for plot
    x = "Year",
    y = "Life Expectancy (in years)",
    color = "Country"
  ) +
  theme_classic() #change theme (remove grid lines)
#Vertical bar chart for average total population from 2012-2021
ggplot(
  data = countryAverageDF %>% filter(Country %in% c("Chad", "Lesotho", "Nigeria", "Japan", "
  mapping = aes(
    x = reorder(Country, -Total_Population), #reorder bars to be descending from highest to
    y = Total_Population)
) +
  geom_col(fill="darkblue")+
  geom_text(aes(label = round(Total_Population)), nudge_y = 10000000) + #add respective valu
  labs(
    x = "Country",
    y = "Average Total Population (to Nearest Person)"
  ) +
  theme_classic()

#Vertical bar chart for average GDP per capita from 2012-2021
ggplot(
  data = countryAverageDF %>% filter(Country %in% c("Chad", "Lesotho", "Nigeria", "Japan", "
  mapping = aes(
    x = reorder(Country, -GDP_Per_Capita),
    y = GDP_Per_Capita) #reorder bars to be descending from highest to lowest
) +
  geom_col(fill="gold")+
  geom_text(aes(label = round(GDP_Per_Capita, 2)), nudge_y = 3000) + #add respective values
  labs(
    x = "Country",
    y = "Average GDP per Capita (USD/Person)") +
  theme_classic()
#Sanitary Expense per GDP line chart - low 3
ggplot(
  data = healthData %>% filter(Country %in% c("Chad", "Lesotho", "Nigeria")),
  mapping = aes( #select dimensions to plot
    x = Year,
    y = Sanitary_Expense_Per_GDP,
    color = Country
  )
```

```
) +
  geom_point()+
  geom_line()+
  labs( #select labels for plot
    x = "Year",
    y = "Healthcare Expense per GDP (Percentage)",
    color = "Country"
  ) +
  theme_classic() #change theme (remove grid lines)


#Life expectancy line chart for low 3 countries over time
ggplot(
  data = healthData %>% filter(Country=="Nigeria" | Country=="Lesotho" | Country =="Chad"),
  mapping = aes( #select dimensions to plot
    x = Year,
    y = Life_Expectancy,
    color = Country
  )
) +
  geom_point()+
  geom_line()+
  labs( #select labels for plot
    x = "Year",
    y = "Life Expectancy (in years)",
    color = "Country"
  ) +
  theme_classic() #change theme (remove grid lines)
#Sanitary Expense per GDP line chart - high 3
ggplot(
  data = healthData %>% filter(Country %in% c("Japan", "Spain", "Switzerland")),
  mapping = aes( #select dimensions to plot
    x = Year,
    y = Sanitary_Expense_Per_GDP,
    color = Country
  )
) +
  geom_point()+
  geom_line()+
  labs( #select labels for plot
    x = "Year",
    y = "Healthcare Expense per GDP (Percentage)",
    color = "Country"
  ) +
```

```
  theme_classic() #change theme (remove grid lines)


#Life expectancy line chart for high 3 countries over time
ggplot(
  data = healthData %>% filter(Country=="Japan" | Country=="Spain" | Country =="Switzerland"
  mapping = aes( #select dimensions to plot
    x = Year,
    y = Life_Expectancy,
    color = Country
  )
) +
  geom_point()+
  geom_line()+
  labs( #select labels for plot
    x = "Year",
    y = "Life Expectancy (in years)",
    color = "Country"
  ) +
  theme_classic() #change theme (remove grid lines)
#Air Pollution Levels - all countries
ggplot(
  data = healthData %>% filter(Country %in% c("Chad", "Lesotho", "Nigeria", "Japan", "Spain"
  mapping = aes( #select dimensions to plot
    x = Year,
    y = Air_Pollution,
    color = Country
  )
) +
  geom_point()+
  geom_line()+
  labs( #select labels for plot
    x = "Year",
    y = "Air Pollution (micrograms per cubic meter)",
    color = "Country") +
  theme_classic() #change theme (remove grid lines)

#Safe Water Access Percentage - all countries
ggplot(
  data = healthData %>% filter(Country %in% c("Chad", "Lesotho", "Nigeria", "Japan", "Spain"
  mapping = aes( #select dimensions to plot
    x = Year,
    y = Safe_Water_Access_Percent,
    color = Country
```

```
  )
) +
  geom_point()+
  geom_line()+
  labs( #select labels for plot
    x = "Year",
    y = "Safe Water Access Percentage",
    color = "Country") +
  theme_classic() #change theme (remove grid lines)

#Life expectancy and safe water access percentage line chart for low 3
ggplot(
  data = healthData %>% filter(Country %in% c("Chad", "Lesotho", "Nigeria")),
  mapping = aes( #select dimensions to plot
    x = Year,
    y = Life_Expectancy,
    color = Country
  )
) +
  geom_point(aes(size = Safe_Water_Access_Percent))+
  geom_line()+
  labs( #select labels for plot
    x = "Year",
    y = "Life Expectancy (in Years)",
    color = "Country",
    size = "Safe Water Access Percent"
  ) +
  theme_classic() #change theme (remove grid lines)

#Life expectancy and safe water access percentage line chart for high 3
ggplot(
  data = healthData %>% filter(Country %in% c("Japan", "Spain", "Switzerland")),
  mapping = aes( #select dimensions to plot
    x = Year,
    y = Life_Expectancy,
    color = Country
  )
) +
  geom_point(aes(size = Safe_Water_Access_Percent))+
  geom_line()+
  labs( #select labels for plot
    x = "Year",
    y = "Life Expectancy (in Years)",
    color = "Country",
```

```
    size = "Safe Water Access Percent"
  ) +
  theme_classic() #change theme (remove grid lines)
```

References

Galasso, M. (2024). Global Health and Development (2012-2021). *Kaggle*. https://www.kaggle.com/datasets/martinagalasso/global-health-and-development-2012-2021

United Nations Secretariat, S. D. of the. (1999). *Standard country or area codes for statistical use*. https://unstats.un.org/unsd/methodology/m49/overview/

Wang, H., Paulson, K. R., Pease, S. A., Watson, S., Comfort, H., Zheng, P., Aravkin, A. Y., Bisignano, C., Barber, R. M., Alam, T., Fuller, J. E., May, E. A., Jones, D. P., Frisch, M. E., Abbafati, C., Adolph, C., Allorant, A., Amlag, J. O., Bang-Jensen, B., … Murray, C. J. L. (2022). Estimating excess mortality due to the COVID-19 pandemic: a systematic analysis of COVID-19-related mortality, 2020–21. *The Lancet*, *399*(10334), 1513–1536. https://doi.org/10.1016/s0140-6736(21)02796-3