I. Introduction

In this project, our team: Ethan Tebbe, Stephen Ferrier, Ayannah Clouden, Sarah Ruth, Spencer Garrett delved into what life expectancy is impacted by and which factors contribute both positively and negatively to life expectancy in different countries around the world. We decided to use the data set “Country Life Expectancy” due to the data set having enough information (1365 rows X 19 columns with 18 numeric columns) to work with while not going past our timeline. That dataset can be found here: https://www.kaggle.com/datasets/pedramabdolahi/country-life-expectancy

The primary colors used for most of the project can be found here: https://coolors.co/09f574-c16200-881600-4e0110-4a7c59.

Throughout this project each team member came up with their own hypothesis about the different factors of life expectancy and individually followed through with their research. Each team member's findings will be laid out individually to reflect their own thoughts and process during this data investigation. Our motivation behind choosing this topic stems from an understanding that while there are several factors outside of our control, having information on life expectancy would benefit more than just our team. Our primary objective was to find the relationship between the different contributing factors and determine their impact on life expectancy.

Definition of Life Expectancy

> “Life expectancy, estimate of the average number of additional years that a person of a given age can expect to live. The most common measure of life expectancy is life expectancy at birth. Life expectancy is a hypothetical measure. It assumes that the age-specific death rates for the year in question will apply throughout the lifetime of individuals born in that year. The estimate, in effect, projects the age-specific mortality (death) rates for a given period over the entire lifetime of the population born (or alive) during that time. The measure differs considerably by sex, age, race, and geographic location. Therefore, life expectancy is commonly given for specific categories, rather than for the population in general” (Bezy, 1).

II. Data Cleaning

Our first goal with our dataset was to make sure that we had prepared and cleaned it for later use. This process began with importing the csv file into a script using pandas. It was imported to a dataframe, and we used a few functions to figure out what we needed to change about the dataset before it would be usable.

Foremost, we checked if the dataset contained missing or null values in the rows. We were pleased to find out that all of the rows of the dataframe were intact; nothing was missing from our dataset.

Second, we made sure that the data types of all columns matched to the type of data stored in that column. We found that the columns in the dataframe matched to the type of data being stored there across the board. All of the numerical data matched to an appropriate typing of integer or float (year, life expectancy, polio incidence, tuberculosis deaths, tuberculosis incidence, malaria deaths, malaria incidence, alcohol deaths, smoking deaths, obesity deaths, cardiovascular disease incidence, cardiovascular disease deaths, deaths by suicide, mean years of schooling, population, GDP, gov health expenditure, and undernourishment), and the singular text column had a string as its data type. We did not need to do any further work with the data types of our dataset.

```
Data columns (total 19 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   Country                           1365 non-null   object
 1   Year                              1365 non-null   int64
 2   Life expectancy                   1365 non-null   float64
 3   Polio incidence                   1365 non-null   int64
 4   Tuberculosis deaths               1365 non-null   float64
 5   Tuberculosis incidence            1365 non-null   float64
 6   Malaria deaths                    1365 non-null   int64
 7   Malaria incidence                 1365 non-null   float64
 8   Alcohol deaths                    1365 non-null   float64
 9   Smoking deaths                    1365 non-null   float64
 10  Obesity deaths                    1365 non-null   float64
 11  Cardiovascular disease incidence  1365 non-null   float64
 12  Cardiovascular disease deaths     1365 non-null   float64
 13  Deaths by suicide                 1365 non-null   float64
 14  Mean years of schooling           1365 non-null   float64
 15  Population                        1365 non-null   int64
 16  GDP                               1365 non-null   float64
 17  Gov health expenditure            1365 non-null   float64
 18  Undernourishment                  1365 non-null   float64
dtypes: float64(14), int64(4), object(1)
memory usage: 202.7+ KB
```

Third, we had to downsize the amount of columns in our dataframe. Within our project proposal, we had decided the areas that were to be explored within the dataset. These areas did not include a majority of the data included within. In order to save a little bit of time, a little bit of memory, and make our data a lot easier to use, we decided to cut out nine of the nineteen columns from the dataframe. These columns were: *polio incidence, tuberculosis deaths, tuberculosis incidence, malaria deaths, malaria incidence, alcohol deaths, cardiovascular disease incidence, cardiovascular disease deaths, and undernourishment.* After this change, our dataframe contained ten columns and we were prepared to move on to the next step.

For our final step, we decided to look at the naming conventions for the columns. Each column was named in title casing (at least for the first word) with spaces between each word. While this sort of naming convention works well when one wants their data to be legible, the final decision was to rename the columns so that it would be easier to use during the coding portion of our exploratory data analysis. We made all letters lowercase and replaced spaces with underscores using the following code:

```
clean_df.columns = [x.lower().replace(' ', '_') for x in clean_df.columns
```

This returned our columns to us in a usable state; for example, '*mean years of schooling*' became '*mean_years_of_schooling*'. While these column names were still a tad long for proper easy-to-write coding names, our group did not feel the need to shorten the lengthier names to save a few keystrokes down the line.

After everything that needed to change about our dataset had been changed, we exported the clean data into a new csv file titled *clean_life_expectancy.csv* and we began the next steps of the project.

III. **Research Questions**

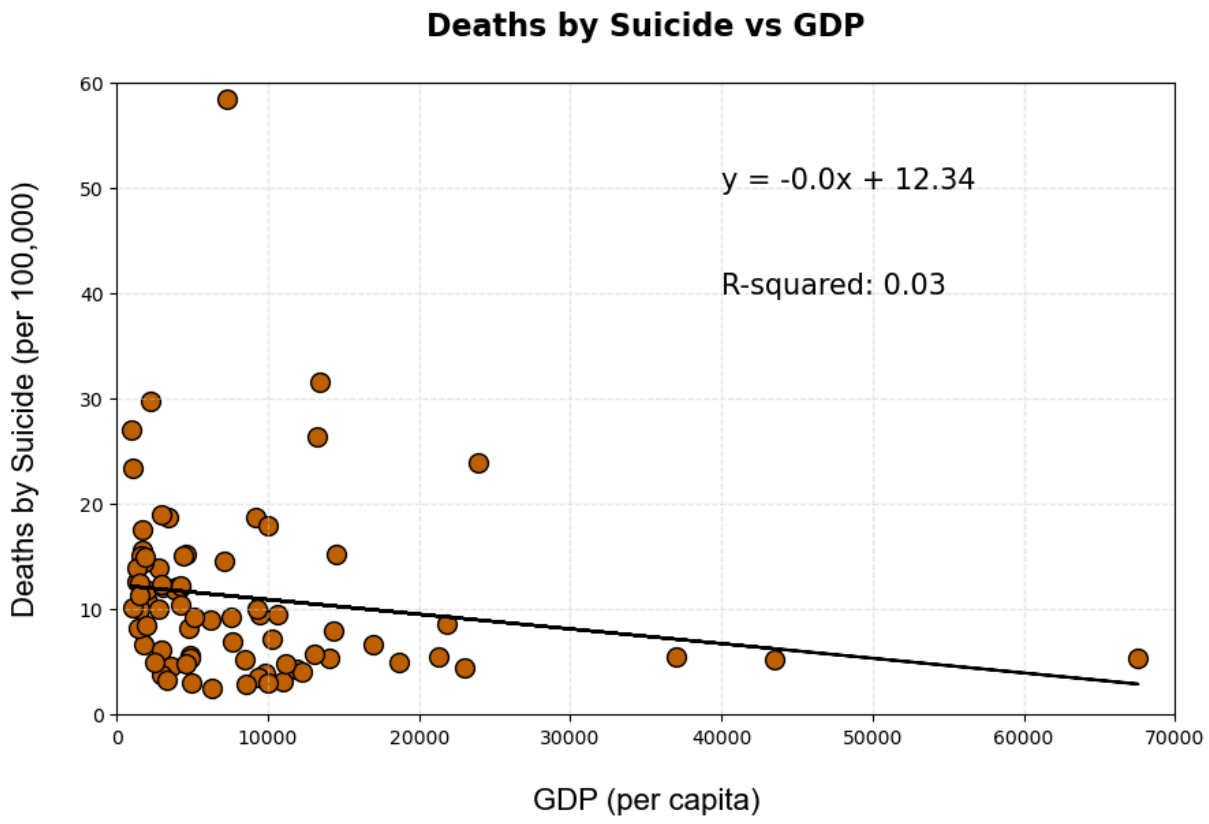
Is there a relationship between deaths by suicide and GDP?
Is there a relationship between deaths by suicide and mean years of schooling?

The Life Expectancy data set contained many variables, but I was particularly interested in the data regarding suicide. I wanted to know if there was a relationship between deaths by suicide and GDP as well as deaths by suicide and mean years of schooling. I began my analysis in my jupyter notebook by importing my dependencies which included: pandas, numpy, hvplot.pandas, matplotlib.pyplot, seaborn, and linregress from scipy.stats. Then I loaded in the csv file we created and saved it to a dataframe. The numeric data on suicide was listed as per 100,000 deaths, the GDP was per capita and the mean years of schooling was the average number of years adults over 25 years old participated in formal education. I grouped the data by country and then found the mean of the variables.

```
df1 = df.groupby('country').agg({'deaths_by_suicide': 'mean', 'gdp': 'mean'}).reset_index()
df2 = df.groupby('country').agg({'deaths_by_suicide': 'mean', 'mean_years_of_schooling':
'mean'}).reset_index()
```

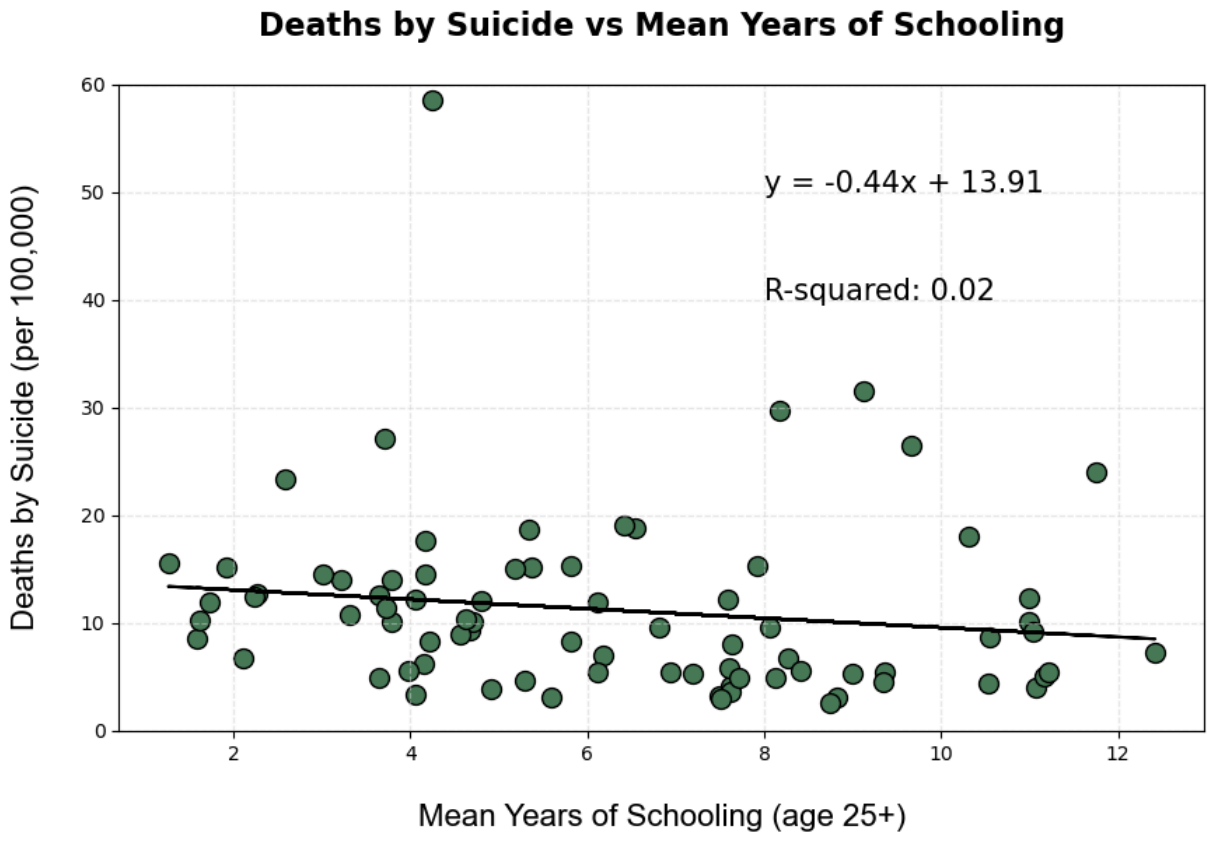
The majority of the code that I used to analyze my research questions came from the activities that were completed in class. I also utilized the expert learning assistant in Canvas as well as additional help from my team.

When looking at suicide and GDP, I started out using a scatter plot and a bubble chart, but then decided that a linear regression would provide more information. This graph and equation shows that there is no relationship. The R-squared value is .03. “In a study by Meda et al, there was no evidence that GDP plays a role in suicide, however it does reveal that increases in unemployment rates are associated with higher rates of death by suicide in males as well as working age males and females” (Meda et al.).



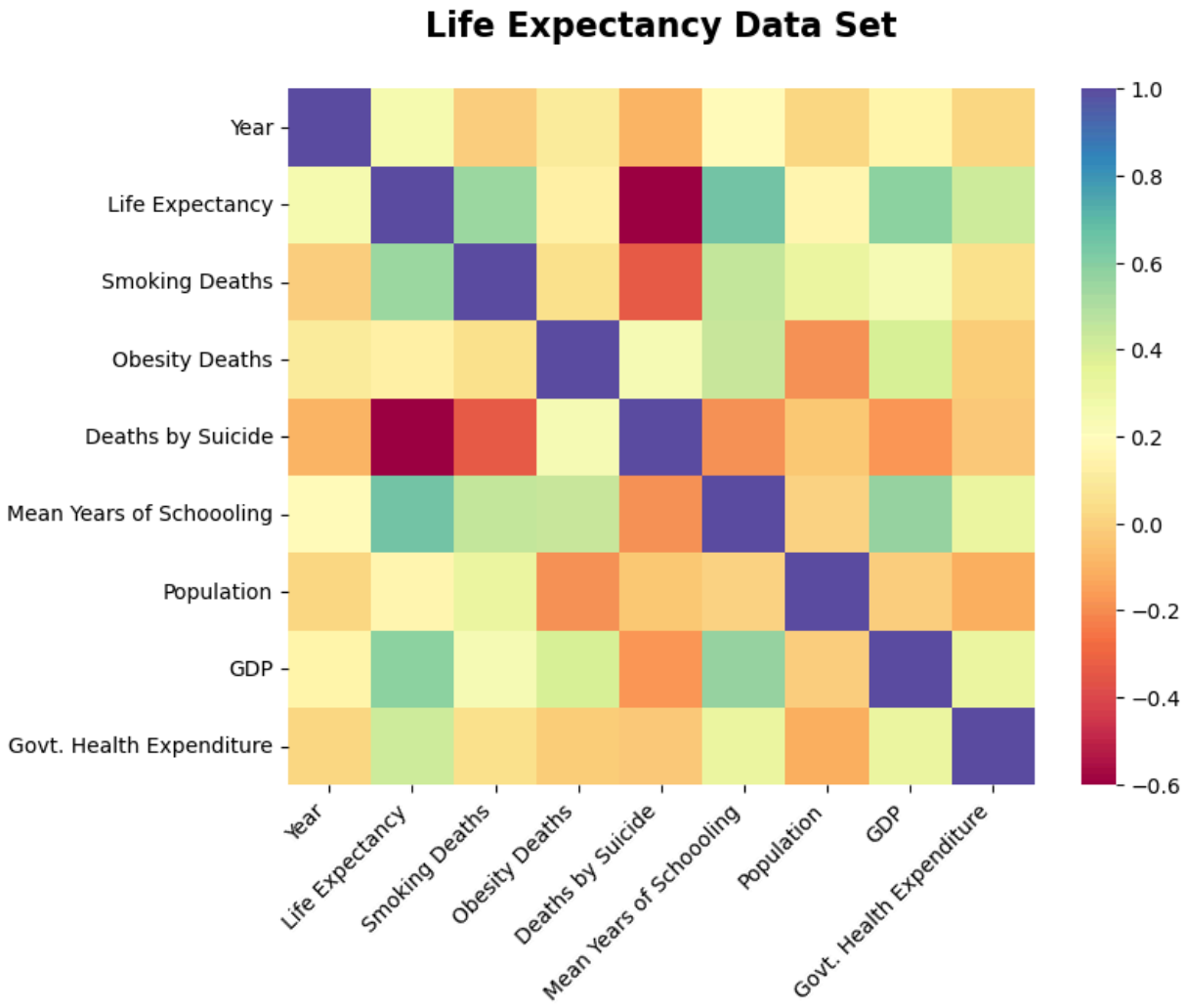
When looking at Deaths by suicide and the mean years of schooling, there also was no relationship. The R squared value is .02. Although there is not a relationship, schooling is most likely a confounding factor.

“In a study by Khazaei et al., the suicide rate varies greatly between countries with different development levels. Our findings also suggest that male gender and human development Index (HDI) components are associated with increased risk of suicide behaviors” (Khazaei et al.).

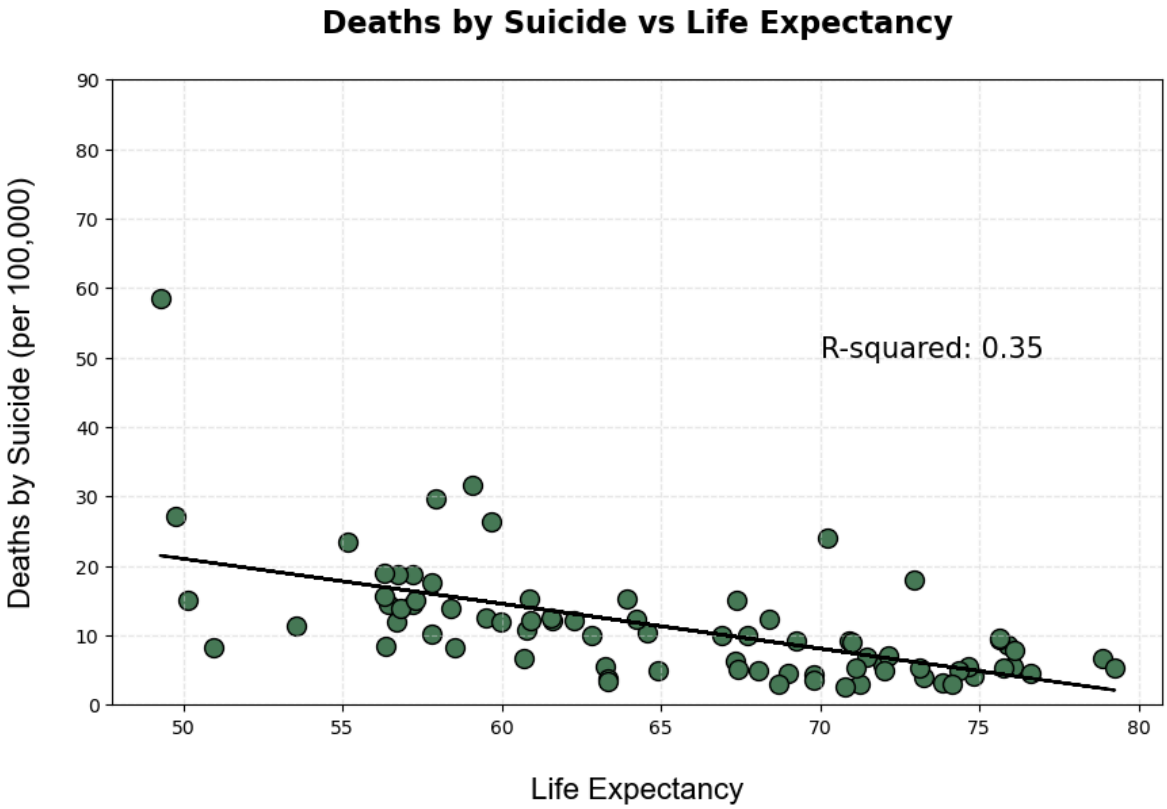


When looking at the deaths by suicide vs. GDP and deaths by suicide vs mean years of schooling, there were two main outliers. The country of Eswatini (formerly Swaziland) had a the highest deaths by suicide at 57 per 100,000 deaths." According to the CIA website, despite its classification as a lower-middle income country, Eswatini suffers from severe poverty, corruption and high unemployment" ("The World Factbook 2021"). All of these factors as well as additional unnamed factors contribute to this data. The United Arab Emirates had the highest GDP at $71,782 per capita. "Strategic location, strong financial reserves, large sovereign wealth fund, promising investor home economies, consistent government spending, progressive policy of economic diversification, free zones and increased foreign direct investment contribute to the UAE's robust economy" ("About the UAE Economy") .

After I finished the two linear regressions, I decided to take another look at all of the variables that are in our clean dataset by creating a correlation heatmap.



The heatmap indicated there may be a negative correlation between deaths by suicide and life expectancy, so I wanted to inquire further using another linear regression.



I did find that there is a weak negative correlation, with a R squared value of .35. The correlation gets stronger the closer the R-squared value is to 1. According to ChatGPT, this could suggest that efforts to improve healthcare, support systems and overall well-being are making a positive impact on society.

Although the correlation heat map on all the variables are the last graphs that I created, the team thought that this data and visualizations would make sense to go to the beginning of our presentation.

Is there any correlation between a country's obesity death rate and their GDP?

After reviewing the Life Expectancy dataset, I became curious about the relationship between GDP and obesity death rates. I wanted to know if a higher or lower GDP value had any correlation between the amount of obesity deaths that specific country had. From this, I formulated this question: Is there any correlation between a country's GDP and their obesity death rate?

Firstly, we should define GDP. An acronym for Gross Domestic Product, GDP is defined by data.oecd.org as:

> *"[...]the standard measure of the value added created through the production of goods and services in a country during a certain period. As such, it also measures the income earned from that production, or the total amount spent on final goods and services (less imports)."*

This same source goes on to say, "While GDP is the single most important indicator to capture economic activity, it falls short of providing a suitable measure of people's material well-being for which alternative indicators may be more appropriate." Though using GDP as a measurement of economic success may not always be accurate, for the research question in mind, it is a great indicator.

I began my analysis by creating a Jupyter Notebook and importing my tools in. For this question, I imported pandas, seaborn, matplotlib, and numpy. I also used linregress and ttest_ind from scipy.stats, and stats from scipy. After reading in the file, I launched a few basic commands just to get a simple view of the data. I explored the df.head, df.columns, and df.dtypes of the dataset.

Then, because the dataset holds multiple rows of data for each country, spanning across years, I gathered the overall mean for each country's obesity death rates and GDP value. This was done and evaluated using this line of code:

```
df1 = df.groupby('country').agg({'obesity_deaths': 'mean',
                                 'gdp': 'mean'}).reset_index()
df1.describe()
```

Next, I determined that a correlation matrix would be the best way to initially determine if there was any type of relationship between the two specific data subsets of obesity death rate and GDP. Before I could conduct the matrix, however, I had to convert the country column from an object to a float. This was done using these two lines of code:

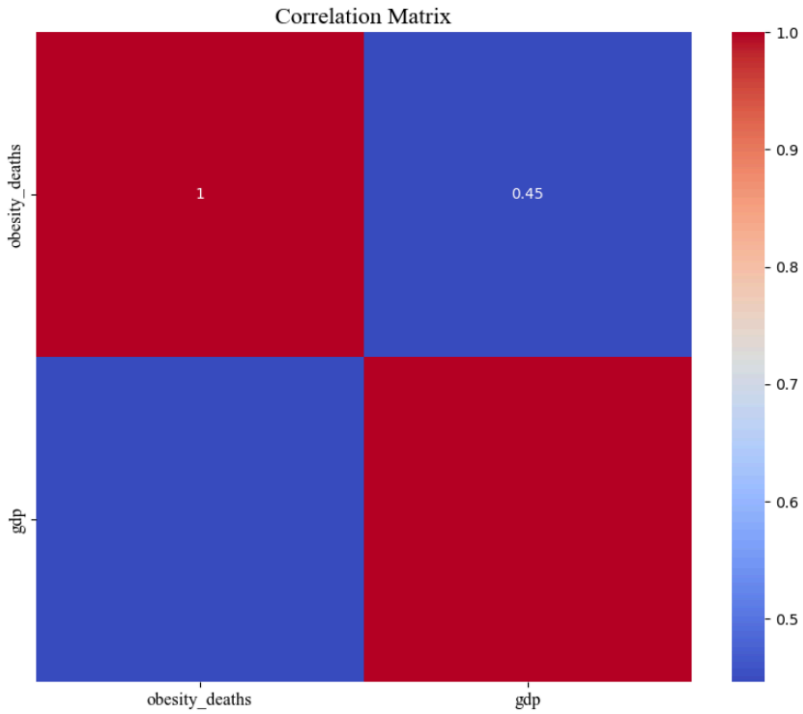
```
df1['country'] = pd.to_numeric(df1['country'], errors='coerce')
df1['country'] = df1['country'].astype(float)
```

Now, I was able to run the matrix. Once it was run, I created a heatmap to visualize the results. The empty country column was visually unappealing in the heatmap, so I added in a line of code to remove it. This line was used to create the correlation matrix:

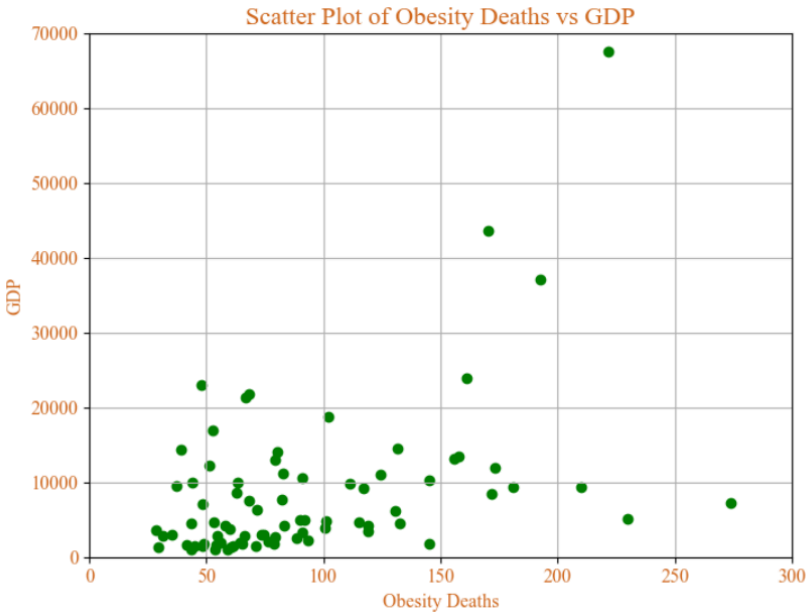
```
correlation_matrix = df1.corr()
print(correlation_matrix)
```

This line was used to remove the empty country column:

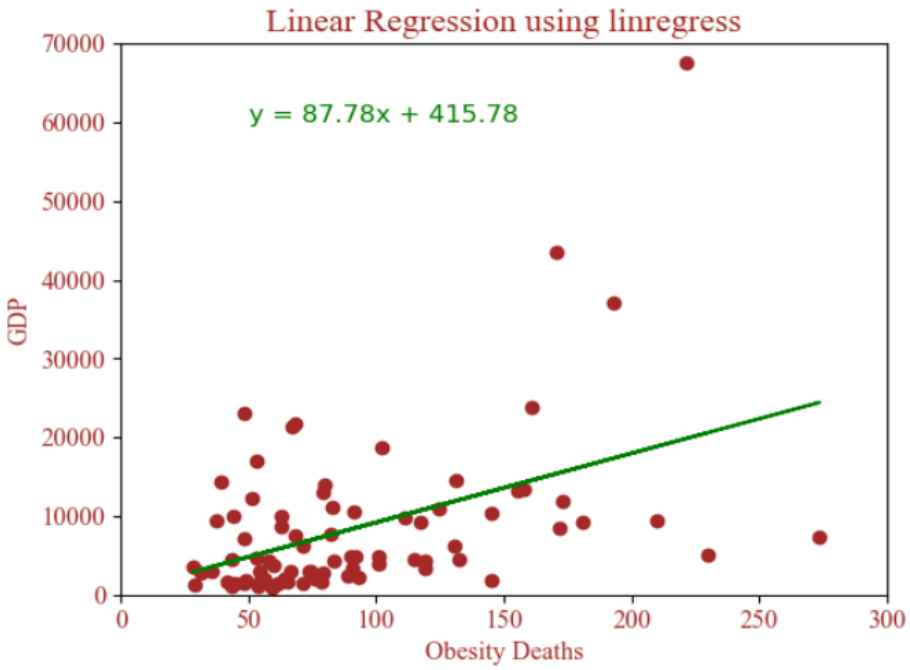
```
correlation_matrix = df1[['obesity_deaths','gdp']].corr()
print(correlation_matrix)
```



With a correlation coefficient of 0.4460823511829229, I determined there to be between a weak to moderate correlation that needed further examination. So I continued my analysis. I decided next to use a scatter plot. It would show a clear relationship between the two subsets of data, and I believed I would have my answer.



However, as we can see, the results of the scatter plot are not particularly strong. There is indeed a weak positive correlation, but as the results of the correlation matrix were also weak, I felt this was still not enough to determine if there was any statistically significant correlation between a country's GDP and the amount of obesity related deaths that country had. So, She Persisted.




```
Slope: 87.77929964287213
Intercept: 415.7755527167001
R-squared: 0.1989894640368845
P-value: 3.784889178986232e-05
Standard Error: 20.070159412582765
```

I decided a linear regression would be able to help me further understand how strong the correlation was between my two subjects. Due to a few extreme values, I believe that the linear regression initially alludes to a stronger correlation than there actually is. With an R-squared of 0.199 and a P-value of 3.78, I was still drawing a conclusion of a positive but weak correlation. With three graphs that all showed weak positive correlation, I decided to attempt to find outside literature on the topic.

What I found upon further reading, was that there was little research done on correlation between obesity *death rates* and GDP. Most of the research available speaks to details and relationships of the *living* obese population and GDP.

Additionally, much of the present material seemed unsure or even conflicting in the information it revealed about obesity and GDP. A study titled, "*Relationship Between National Economic Development and Body Mass Index in Chinese Children and Adolescents Aged 5–19 From 1986 to 2019*" published in 2021 claimed that, "No studies have evaluated how economic expansion impacts the prevalence of obesity." In this study, Bu et al conclude that, "Results from this study show for the first time that higher BMI in both boys and girls and in all age groups between 5 and 19 was strongly related to the economic expansion in the past 35 years in China."

A separate article, published in 2019, titled "What is driving global obesity trends? Globalization or "modernization"?" found in their research that, "There was also evidence of a curvilinear relationship between GDP per capita and BMI: among low income countries,

economic growth predicted increases in BMI whereas among high-income countries, higher GDP predicted lower BMI." (Fox, et all, 2019).

Obesity undoubtedly has an effect and cost on any given economy. It can be said that the information unfolded here can be considered an extension of studies already conducted. A positive correlation between GDP and obesity death rates, despite being weak, could potentially be said to be further evidence of economic growth being a factor in predicting rising obesity rates. However, further research would need to be done to conclude with definitive statements.

Is there a relationship between smoking deaths and life expectancy?

One of the areas of interest in the dataset was centered around the question '*Is there a relationship between smoking deaths and life expectancy?'* To begin exploring the answer to this question, there were a few simplification processes that needed to be done.

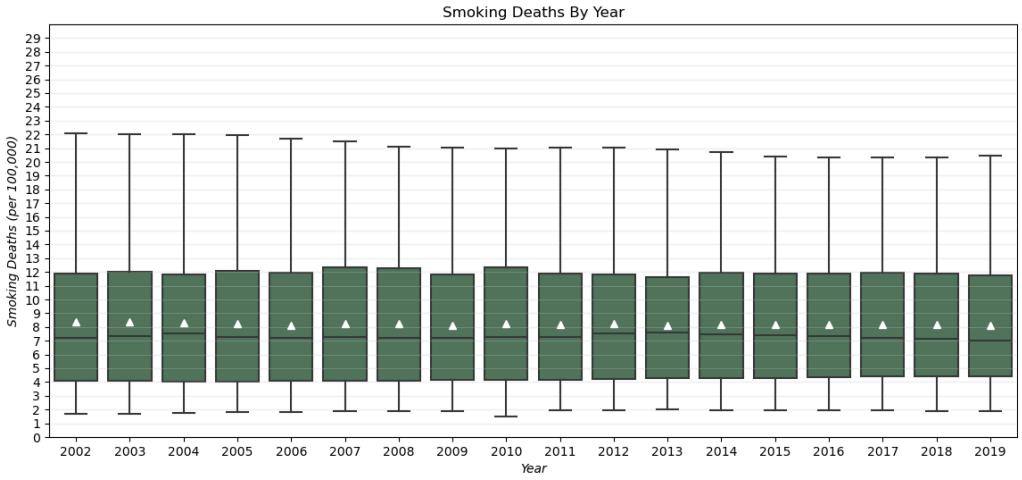
In the clean dataset, there still existed a large number of columns that did not pertain to the stated question. This issue was rectified by cutting down the columns to the bare minimum: *country, year, life_expectancy, smoking_deaths, and population*.

I attempted to make a few cheap visualizations in order to get a better understanding of the data on smoking deaths. However, these visualizations proved to be unusable given the sheer amount of data points being shoved onto the plots. One of these failed visualizations was a line chart showing the change of smoking deaths over time for the seventy-nine countries included in the dataset. This graph was unreadable and uncovered little to no insights about the dataset. In the other attempt, I created a scatter plot showing smoking deaths and life expectancy points with each country as a different color. The linechart was unreadable; however, this scatter plot managed to go another step past that as it contained an egregious amount of entries.

With the failed visualizations tried and tested, it became apparent that a different method was required to explore the question at hand. I ran a set of aggregations on the dataset:

	year			life_expectancy			smoking_deaths			population		
	mean	median	std	mean	median	std	mean	median	std	mean	median	std
country												
Afghanistan	2010.500000	2010.5	5.338539	60.706917	61.13495	2.264490	5.216688	5.288251	0.456762	2.930117e+07	28719414.0	5.102265e+06
Algeria	2010.500000	2010.5	5.338539	73.851694	73.96575	1.717862	10.650280	10.603426	0.218150	3.657804e+07	36199948.0	3.555006e+06
Angola	2010.500000	2010.5	5.338539	56.437317	57.16095	4.711916	5.591321	5.598154	0.094105	2.424106e+07	23811658.0	4.706125e+06
Argentina	2010.500000	2010.5	5.338539	75.902183	76.02995	0.929325	15.199574	15.266941	0.443975	4.134101e+07	41310430.0	2.196984e+06
Armenia	2010.714286	2011.0	5.915151	73.255157	73.30685	1.525564	16.109184	16.099052	0.108813	2.949583e+06	2930362.0	9.815561e+04
...	...	...	...	...	...	...	...	...	...	...	...	...
Uganda	2010.500000	2010.5	5.338539	57.278650	57.50990	4.275366	4.034333	4.084155	0.279457	3.331422e+07	32818731.0	5.366195e+06
United Arab Emirates	2015.500000	2015.5	2.449490	79.253450	79.27895	0.363931	11.391209	11.546614	0.502216	8.948011e+06	8955587.5	1.909207e+05
Uzbekistan	2010.500000	2010.5	5.338539	69.225989	69.44205	1.504054	7.422734	7.375181	0.398605	2.898801e+07	28835845.0	2.350407e+06
Zambia	2010.500000	2010.5	5.338539	56.303300	57.28495	5.296183	5.459506	5.715942	0.446552	1.416387e+07	14028959.5	2.505300e+06
Zimbabwe	2014.500000	2014.5	3.027650	57.924240	59.21855	3.650156	8.401938	8.504893	0.470693	1.403076e+07	14005347.5	8.678907e+05

Behind these aggregations, the idea was to figure out whether the median or mean would better represent the data points–it turns out that either would work fine for the dataset. Both the median and the mean mirrored each other throughout the countries. To be sure, I created a boxplot of the smoking deaths of the entire dataset for each year.



Once again, the mean and the median mirrored each other over the years included in the dataset. Even when considered as a whole over the years included in the dataset, the median and mean remain similar. With that in mind, it made sense to create a new dataframe which condenses each country into a single data point using the average of the smoking deaths.

From the average dataframe, I created a pair of leaderboards to identify the countries with the highest and lowest smoking deaths.

Bottom 5

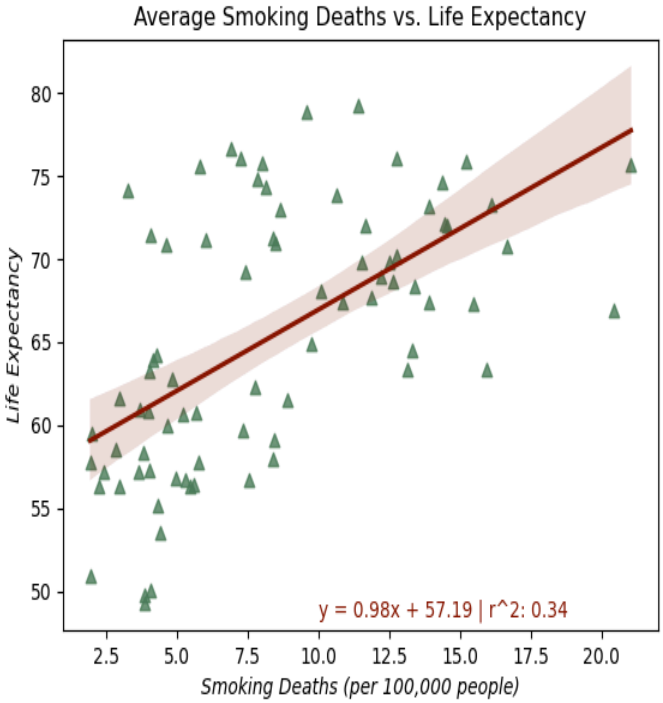
Country	Smoking Deaths (per 100,000 people)
Nigeria	1.94
Niger	1.96
Ethiopia	2.02
Burkina Faso	2.25
Guinea-Bissau	2.44

Top 5

Country	Smoking Deaths (per 100,000 people)
China	21.04
Nepal	20.42
Philippines	16.65
Armenia	16.11
Myanmar	15.95

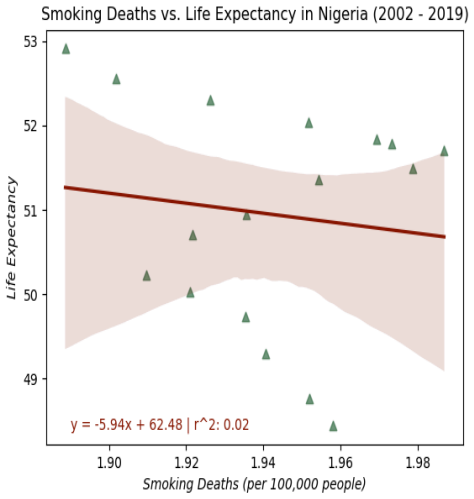
With the insight that the leaderboards provided into the countries with the highest and lowest incidence of smoking deaths, I could take the final step towards finding an answer to my question.

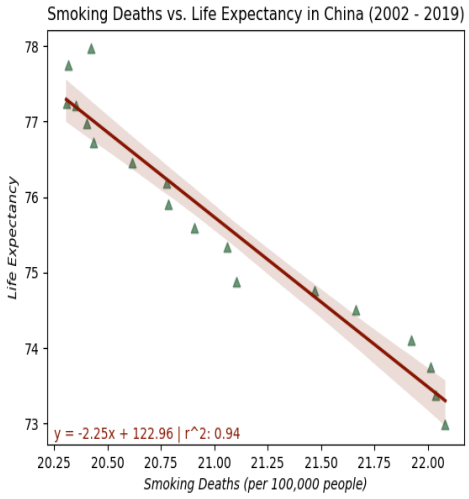
Using the linregress method from the scipy.stats module, I created three linear regression models for the data. One of these focused on the relationship between the average smoking deaths and life expectancy for all countries.



Upon making this regression, it would appear that the relationship between smoking deaths and life expectancy is a little weak but still shows some relationship. To follow up on this finding, I created two more linear regressions: one for the smoking deaths and life expectancy of China (the country with the highest number of smoking deaths) and Nigeria (the country with the lowest number of smoking deaths).

Within the linear regression on China, the relationship between smoking deaths and life expectancy shows a strong positive correlation with an r-squared value of 0.94–this is a promising point in the dataset. However, when compared to the value that is contained in the linear regression on Nigeria, an r-squared value of 0.02, it falls apart.





With the differences between the countries in mind, it would appear that there is no definitive relationship between the smoking deaths and life expectancy of any given country.

To further the discussion of this question, I believe it would have been a great idea to incorporate data points relating to the number of smokers in each country. This data would provide a method to remove hidden outliers–those countries that do not have a large amount of smokers whether it be from lack of access or just lack of people interested in smoking. I believe that issue causes the difference that was observed in the linear regression of Nigeria and it could be affecting more countries in the dataset as well.

What is the relationship between life expectancy and government health expenditure?

I, Spencer, wanted to explore the relationship between life expectancy and government health expenditure to gain a better understanding if there was any at all, and if so what is it?

With a clean dataset to work with I set up my notebook and imported the necessary packages, and read the csv file into a pandas data frame. I then took a closer look at my data of interest:

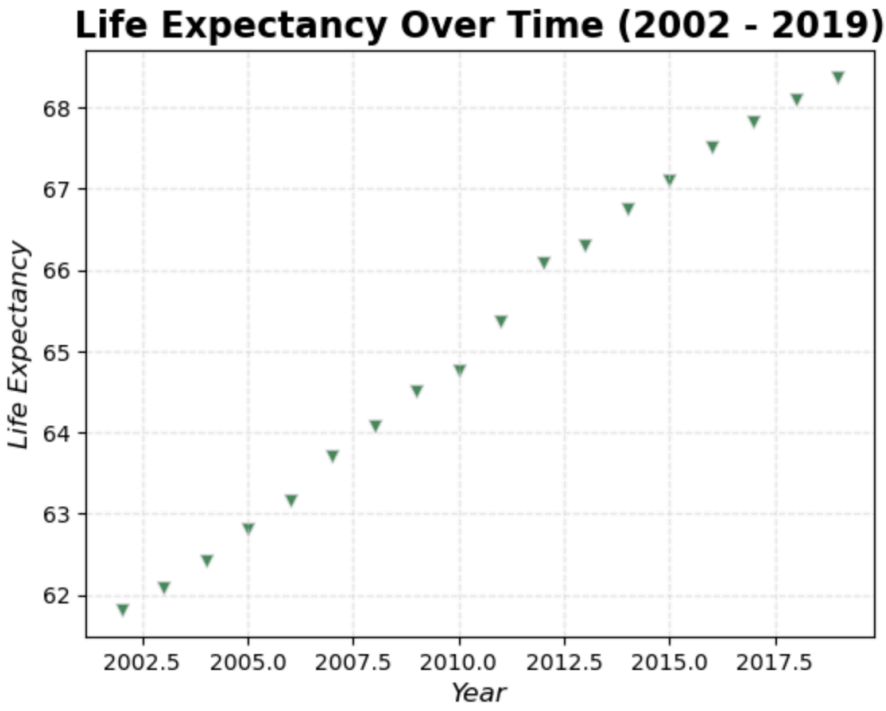
```
cols = ["year","life_expectancy","gov_health_expenditure","country","population"]
df[cols].describe()
```

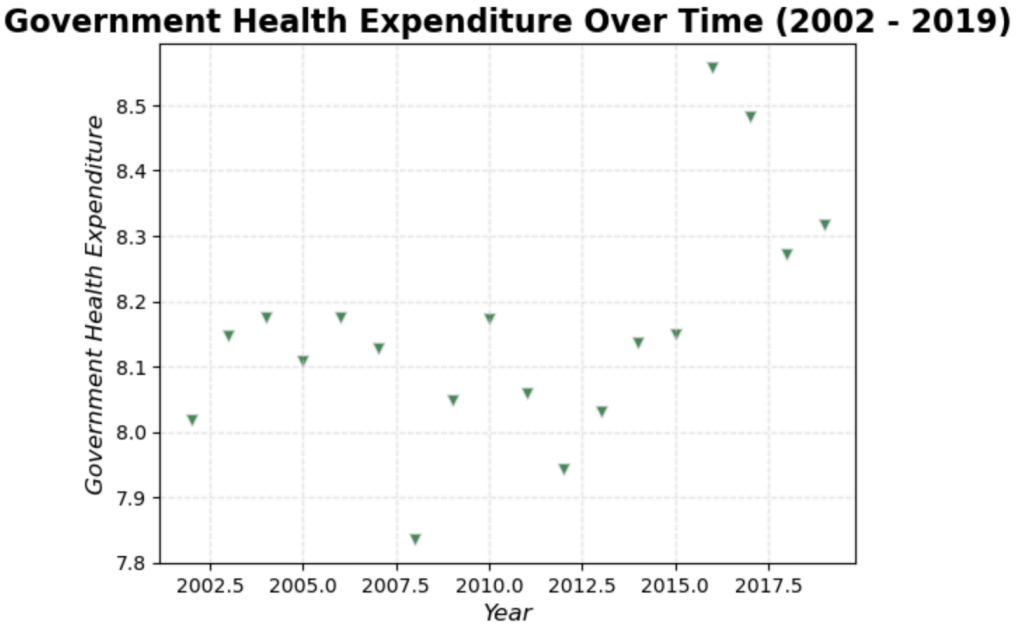
	year	life_expectancy	gov_health_expenditure	population
count	1365.000000	1365.000000	1365.000000	1.365000e+03
mean	2010.675458	65.226921	8.156593	6.636804e+07
std	5.185222	8.134401	4.782531	2.092548e+08
min	2002.000000	42.125400	0.730000	9.717610e+05
25%	2006.000000	59.349100	4.790000	6.044130e+06
50%	2011.000000	65.731400	7.130000	1.523498e+07
75%	2015.000000	71.963800	10.250000	3.510727e+07
max	2019.000000	79.726200	33.100000	1.421864e+09

I included the year, country, and population as extra columns that I thought could add more context to my analysis. As far as my main fields of interest go I could tell that life expectancy did not have any major outlier since the mean and median were fairly close. On the other hand government health expenditure was skewed slightly right with potentially a high outlier since the mean was greater than the median.

It was also around this time that I realized I needed to figure out what the dataset actually meant by government health expenditure, so I went back to the Kaggle page and saw that it was a percentage of general government expenditure. This means that the average government health expenditure is 8.15% of its total budget.

Now that I was a little more comfortable with my data I began to make some basic plots for life expectancy and government health expenditure over time to get more of a feel for how it changed over the years. My first visuals were far too noisy with a point for every country in every year, so I moved to taking the overall average of each year (*df_avg_[life/ghe] = df.groupby("year")["life_expectancy"/"gov_health_expenditure"].mean().reset_index()*), which produced the plots below:

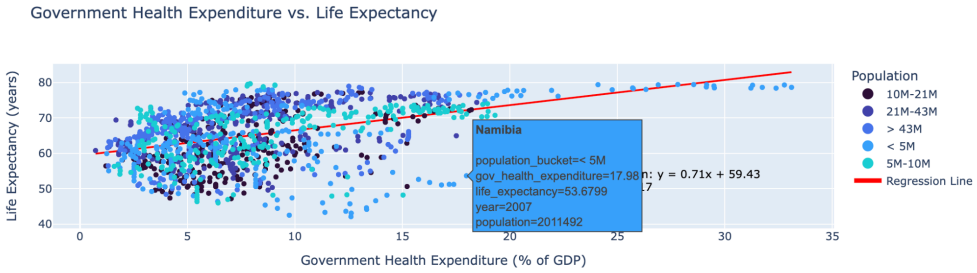




After this I jumped straight into creating a linear regression plot for life expectancy and government health expenditure. I created a new data frame to use for my linear regression work, which worked out well, but I could have used my original data frame since I didn't make any changes at this point. Looking back I also should've done a correlation plot before running my regression, but I didn't circle back to that until later on. The plot below is my first one (after tweaking colors and some design):

I was initially fooled by how positive the line looked with it going up and to the right, but this is caused by the y axis starting at 40 instead of zero. The R-Squared is even more evidence that the relationship is not as direct as it may seem, with a value of 0.17 it is extremely weak. The cluster on the left of data less than or equal to government health expenditure of 20% stood out as well and would be something to explore further by dropping the data points greater than 20%. Relationship aside I also thought that this plot was noisy since it contains a data point for all countries and years.

My first attempt at getting a handle on the noise was to create a plot that colored each point based on country, which resulted in a disastrously long legend for each of the 79 countries included in the dataset. After that I tried to color the plot by year, which looked better since the legend could actually fit in the same space as the plot, but it wasn't clear and easy to understand. I then tried plotting and coloring the dots by population to see if that added more value than country or year, which it did since it was a more manageable five colors. I also ended up trying interactive plots for each one mentioned previously, but my favorite was the population one:

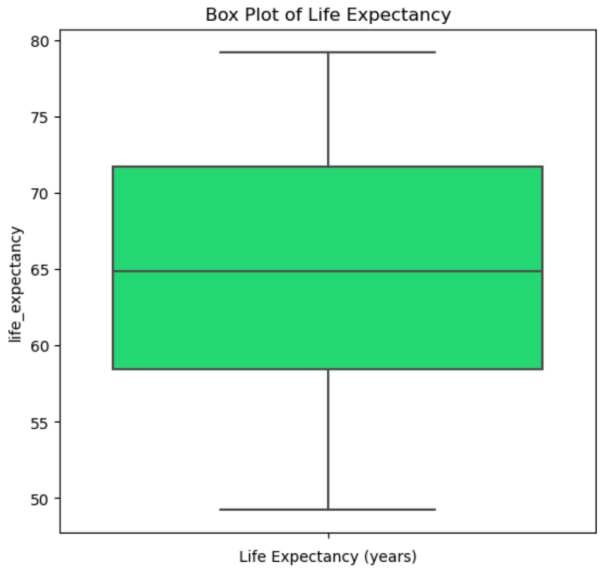


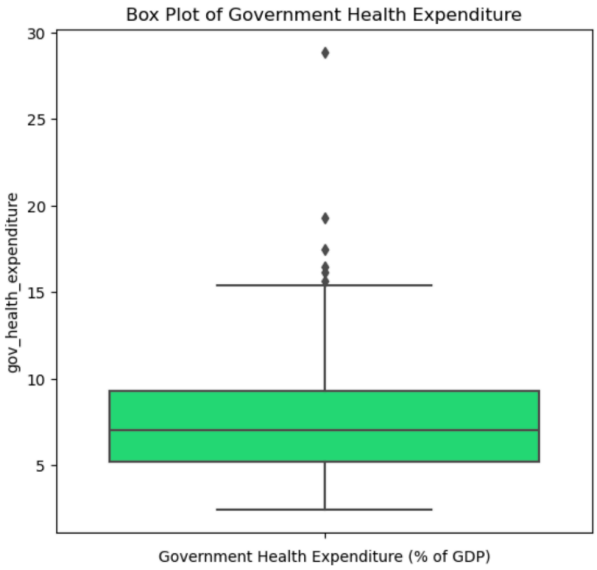
Given how many points the current data frame had I ended up taking the average life expectancy and government health expenditure by year and by country, in respective data frames in order to see which one I wanted to explore further. I stuck with the averages by country since government health expenditure is tied to a country, not a year, and I wanted to be able to see which points correspond to which country:

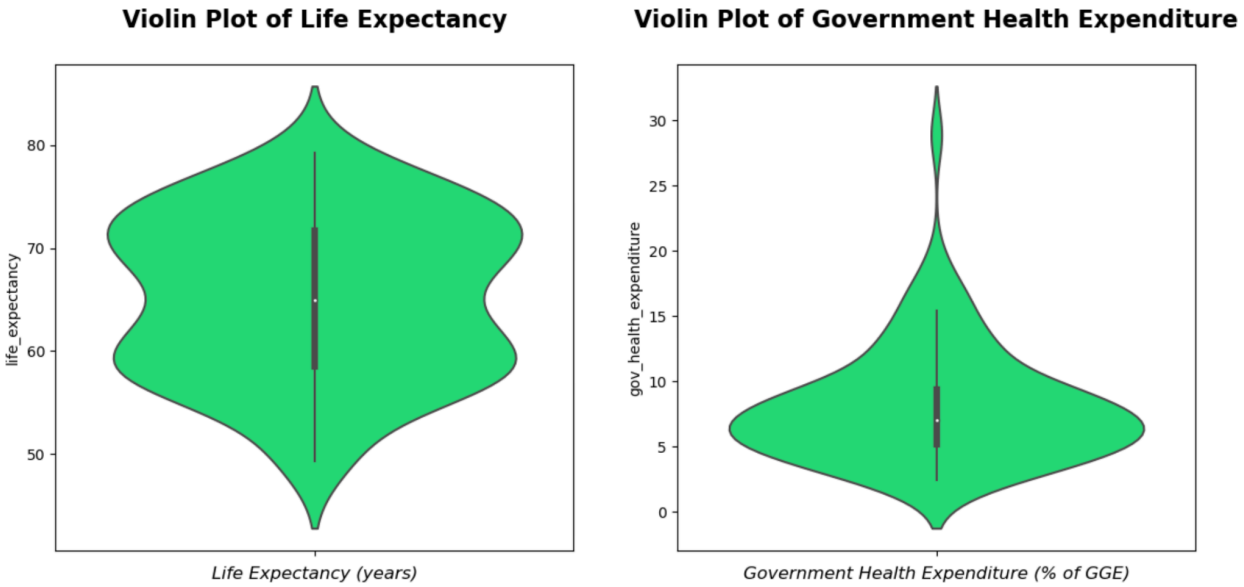
```
# Create dataframe for linear regression with country averages

df_avg_country = df.groupby("country")[["life_expectancy", "gov_health_expenditure", "population"]].mean().reset_index()
df_avg_country.head()
```

	country	life_expectancy	gov_health_expenditure	population
0	Afghanistan	60.706917	2.450000	2.930117e+07
1	Algeria	73.851694	9.350556	3.657804e+07
2	Angola	56.437317	4.747778	2.424106e+07
3	Argentina	75.902183	16.143333	4.134101e+07
4	Armenia	73.255157	6.165714	2.949583e+06

Once I had this average by country data frame to use I began crafting my final visuals to fit our team's color palette and best communicate my findings. I began with box plots and violin plots (opting to use the violin plots in the presentation):

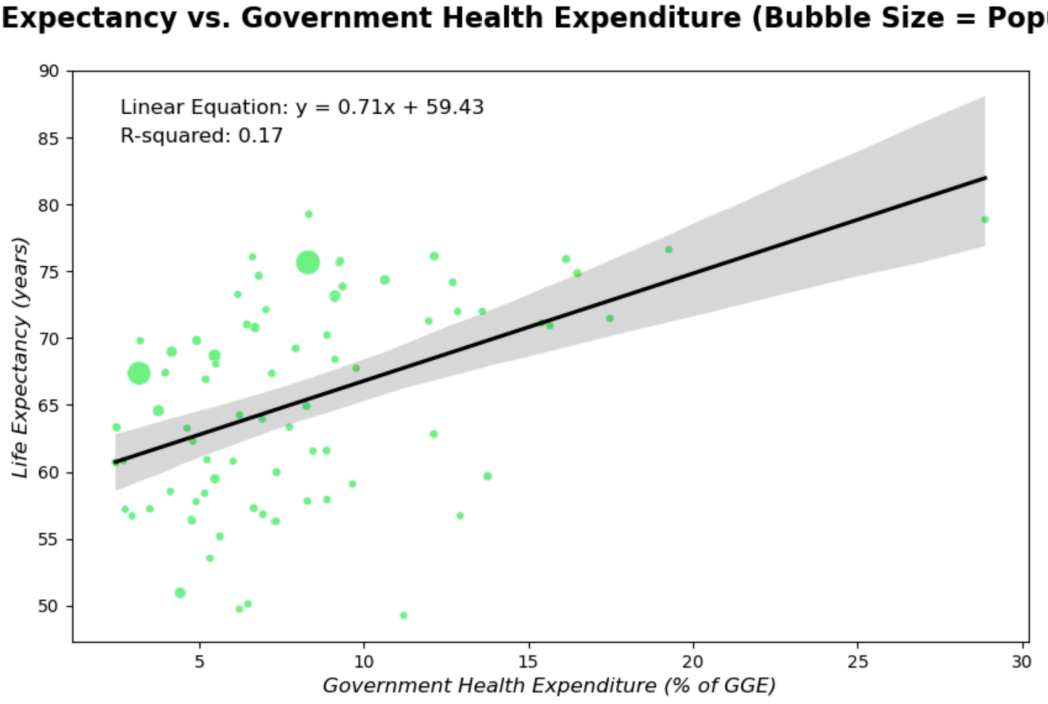






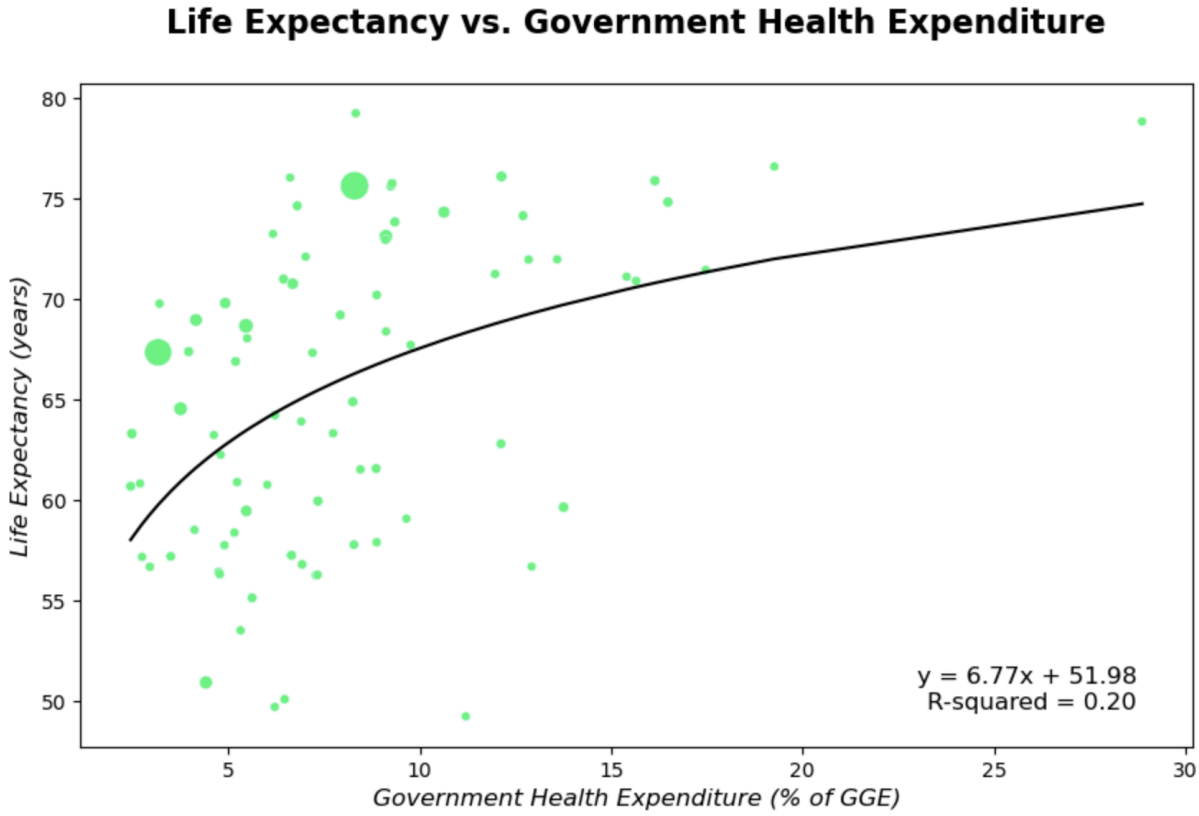
The violin plots display the shape of the data in a much better way than the boxplots. I was able to immediately tell that life expectancy is bimodal with a median around 65, right between the peaks. Government health expenditure told a different story, one with a fairly normal distribution from 0 - 23% and then outliers between 25 and 30%, and a median around 7%.

After seeing each of them individually I went back to recreate the linear regression, and decided to include population to size the dots:



Adding the population to size the dots added more context for myself and served to demonstrate that even a large country (by population) might not spend much on the health of its citizens. Looking back I would now include a legend (hopefully ordered correctly) to show which sizes correspond to which population numbers.

Given how low the r-squared was for the linear regression I thought that maybe a curved line of some sort would fit the data better since it begins clustered on the left and then stays higher on the right. I discovered that a logistic regression is best for data with discrete outcomes, while a logarithmic regression is best for continuous data like mine. So, I gave the logarithmic regression and shot:

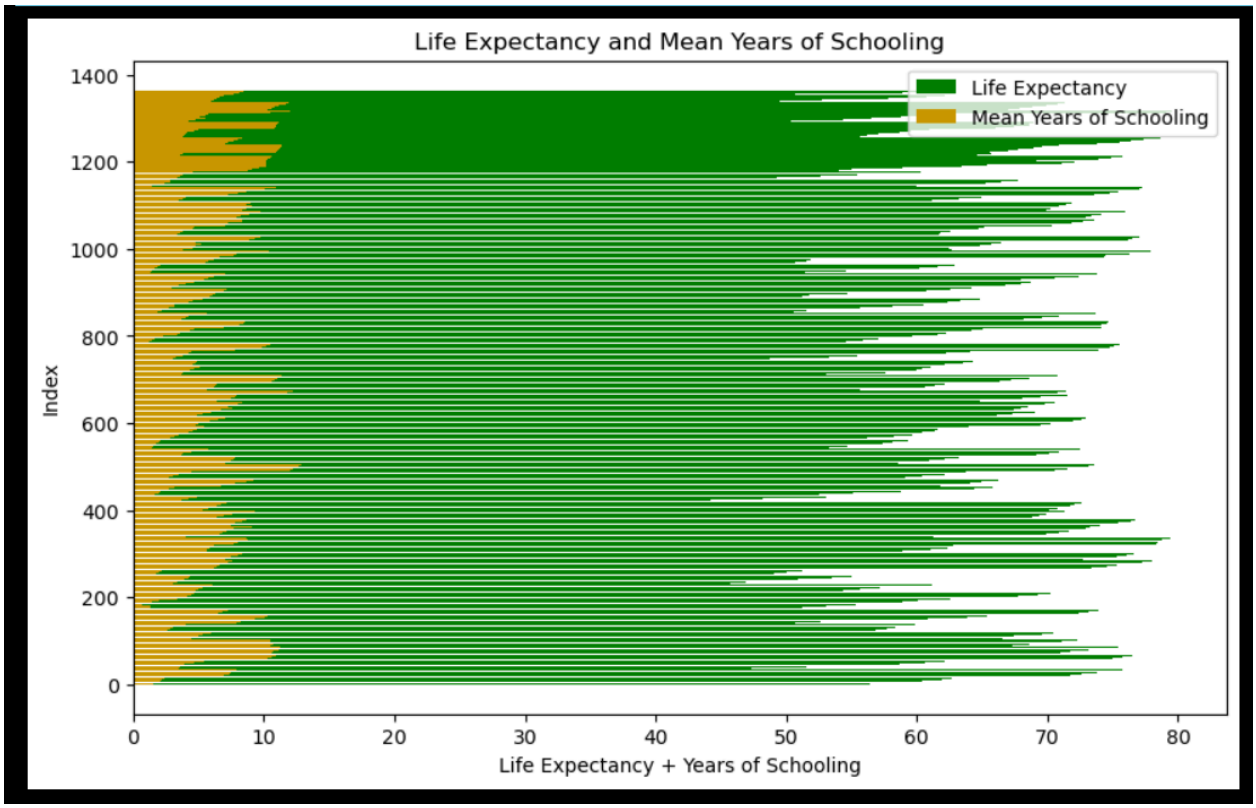


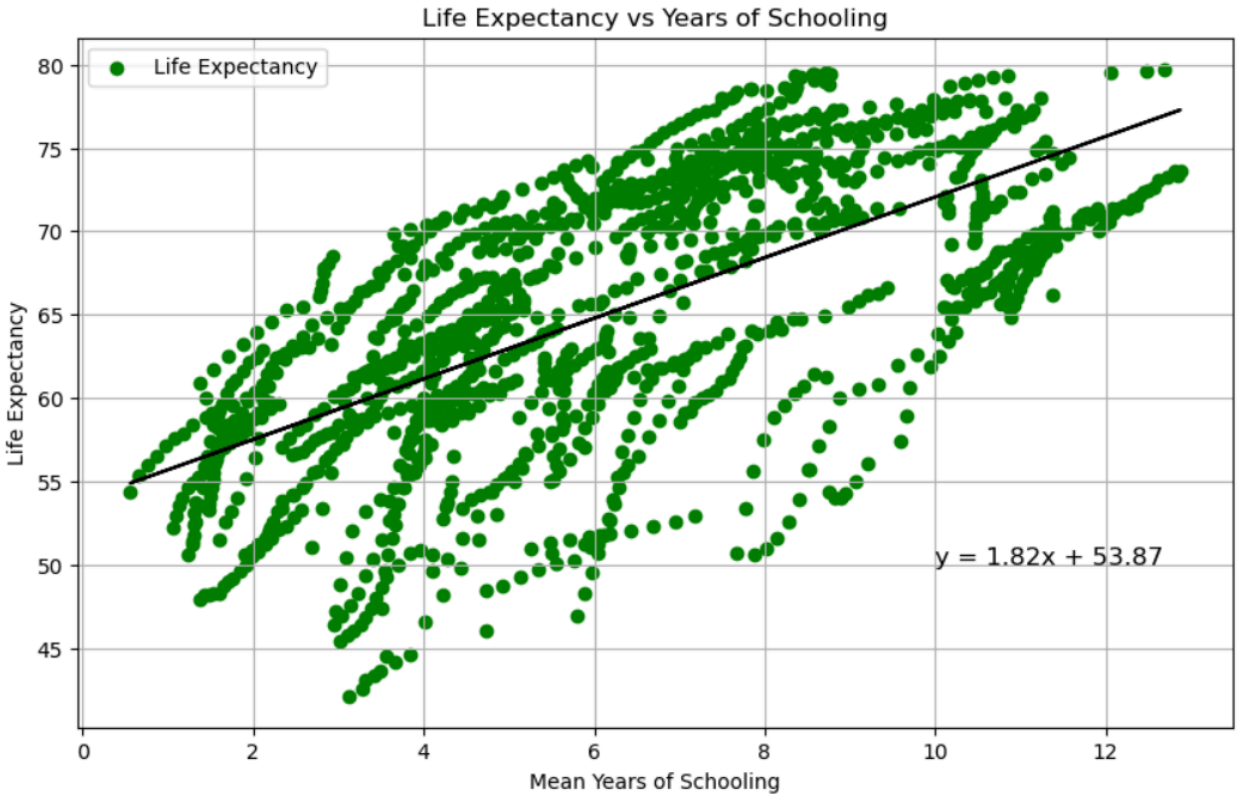
The results were not as great as I expected and only raised the r-squared by 0.03 to 0.20. The cluster on the left does not show much of a pattern with life expectancy varying greatly until health expenditure gets over 15%.

Since the relationship between life expectancy and government health expenditure is so weak there are other factors that should be explored when looking at what can help or hurt life expectancy. This work demonstrates that government health expenditure does not provide the whole picture when looking at life expectancy by country.

Does schooling impact life expectancy?

The schooling section of the study aims to explore whether a longer duration of education correlates with longer life expectancy, and to what extent this is true in various parts of the world. We sourced a comprehensive dataset containing information on 79 different countries on Kaggle.com. After the data cleaning process I took a deeper look into the data set and gravitated to the average years of schooling section. My hypothesis was that more developed countries, which have longer life expectancies on average, would have longer durations of education for the people that live in these countries. The idea was founded on working in the education field for several years and seeing the positive impact education can have on the lives of individuals. With more access to education people can leverage what they know into longer and healthier lives.
To start exploring the data, I compared the life expectancy of each country to their average years of schooling as shown in the charts below:



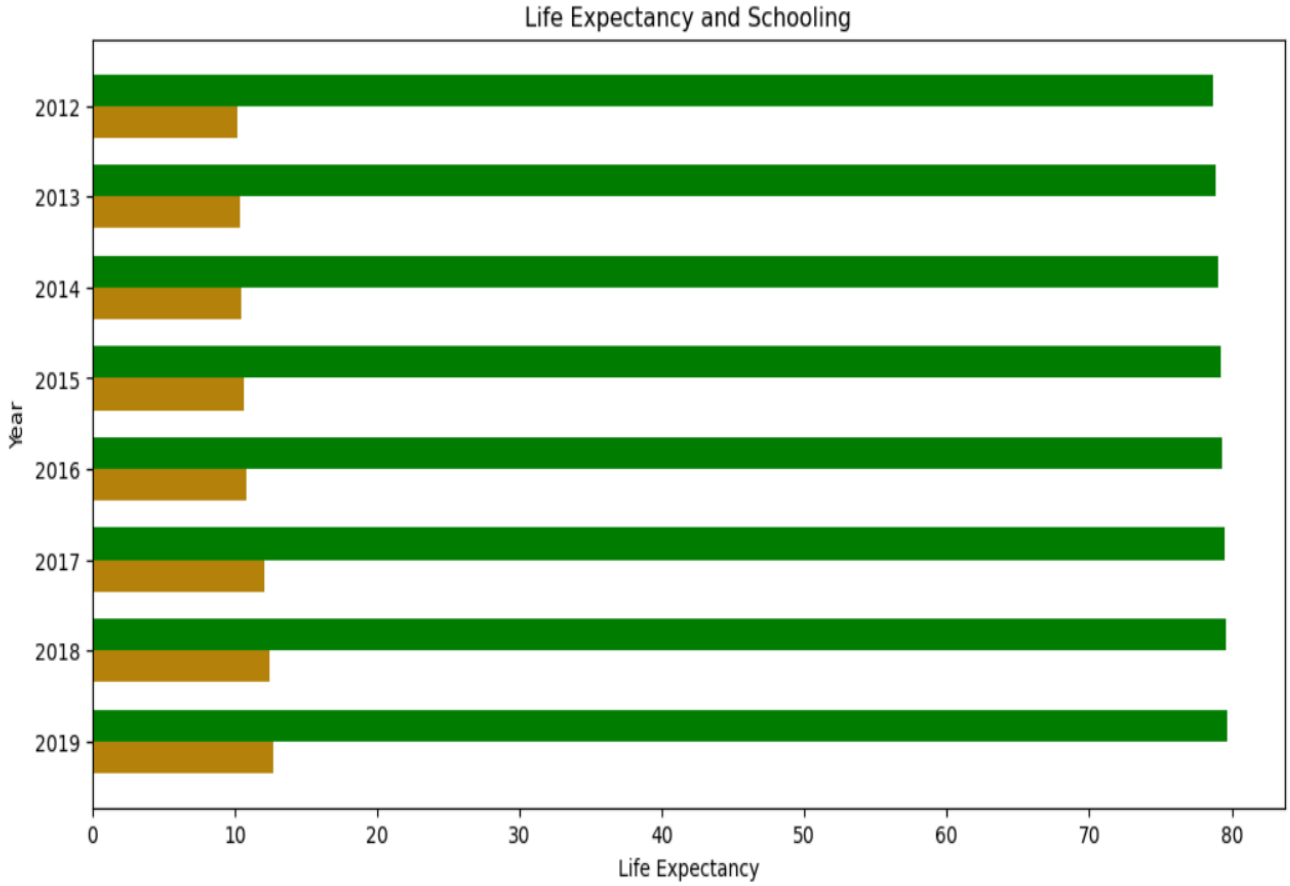


The first chart was an attempt to look at all of the data at once using a horizontal bar chart. Having 79 countries worth of data in a single chart proved to be too much for that style and so I transferred the data to a scatter plot. Overall this plot would support the hypothesis that more schooling results in longer lives. However, this plot does not show any insightful detail and needed to be narrowed down in order to draw any meaningful conclusions. To do that, I took the top five countries and bottom five countries in terms of life expectancy to compare the average years of schooling for just these ten countries.

	Country	Life Expectancy	Mean Years of Schooling
5	United Arab Emirates	79.7000	12.70
6	Costa Rica	79.4000	8.80
7	Thailand	78.9800	8.70
8	Oman	78.0000	11.20
9	China	77.9680	7.60
4	Guinea	59.7199	2.20
3	Mali	59.6600	2.31
2	Central African Republic	55.0253	4.33
1	Chad	53.2600	2.57
0	Nigeria	52.9000	7.20

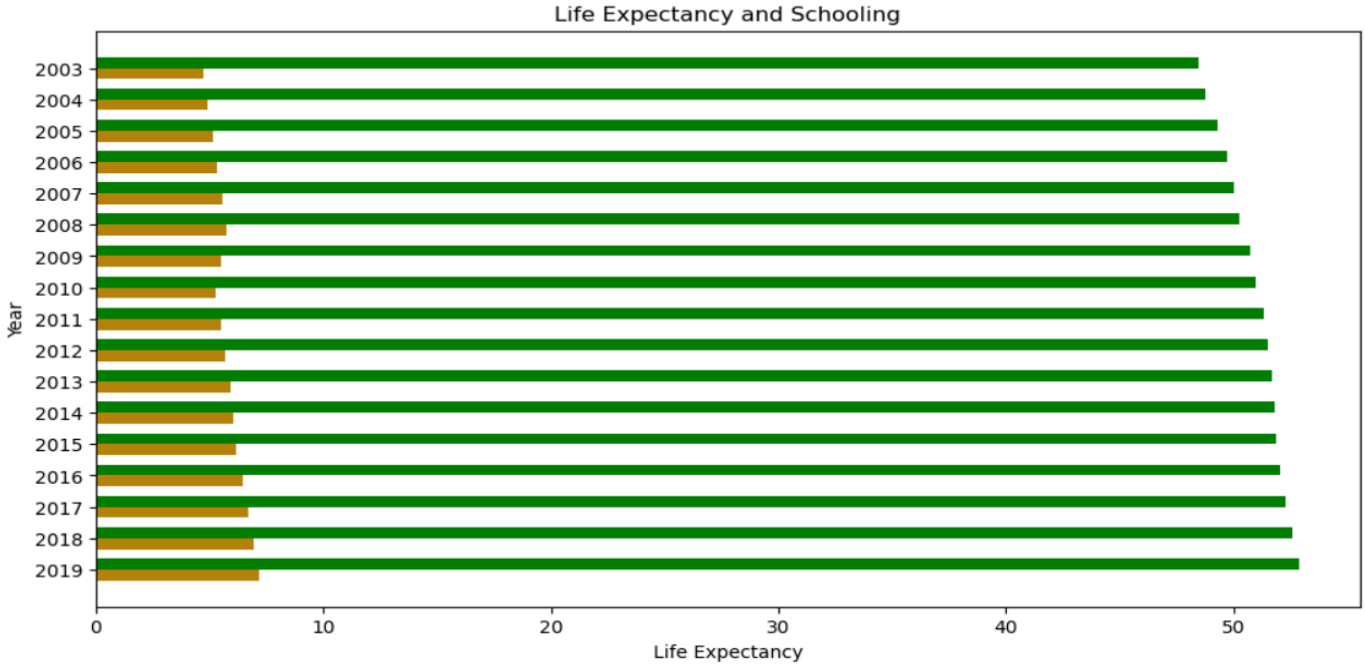
Narrowing down the information being used sped up the process of finding conclusions in the data. Working with these countries, several conclusions were still found after some exploratory research with the available data. I took a look at both the UAE and Nigeria individually as they are the top and bottom countries in my table.

```
Average Life Expectancy in United Arab Emirates: 79.25345000000002
Average Mean Years of Schooling in United Arab Emirates: 11.22061175
```

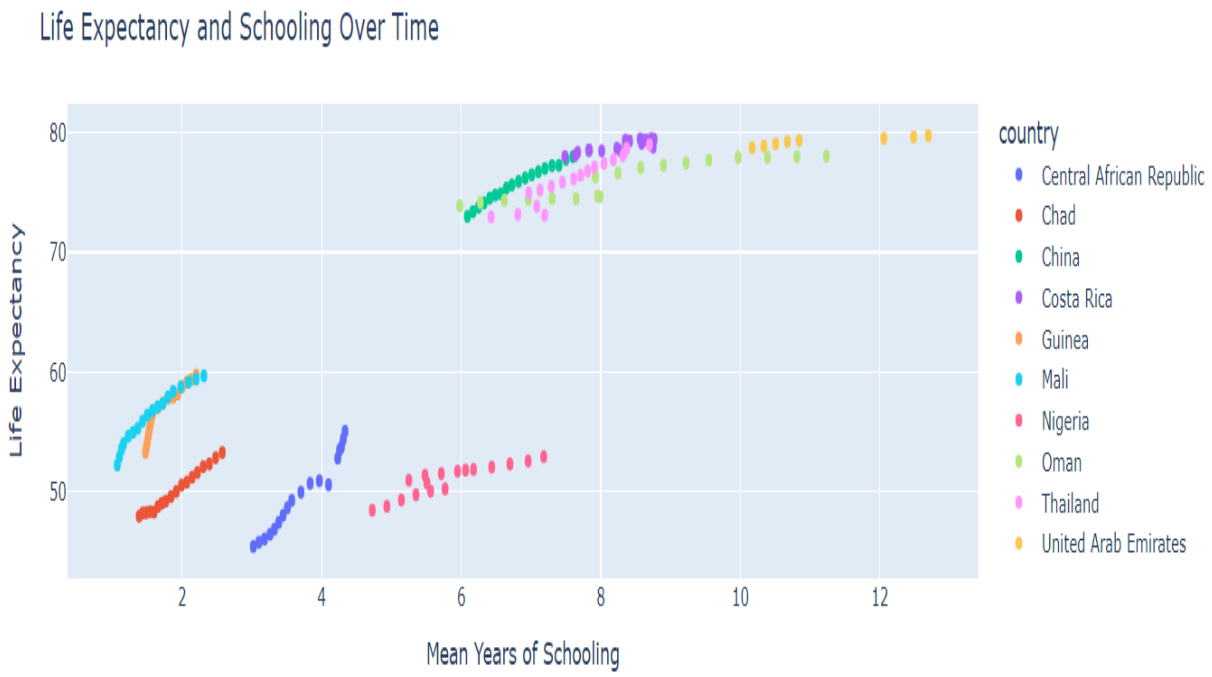


Although this chart shows more detailed data, I began to see the limitations when comparing the UAE to Nigeria.

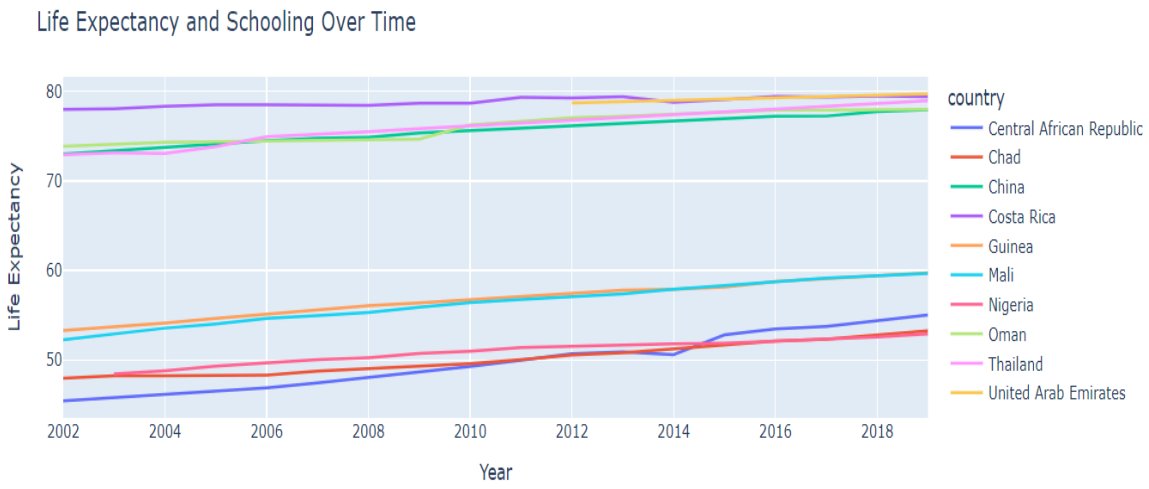
```
Average Life Expectancy in Nigeria: 50.95036470588235
Average Mean Years of Schooling in Nigeria: 5.816890323529412
```



As shown, Nigeria had several more years of data to work with meaning a comparison between the two countries could not reach full conclusion as there is missing data that we do not have access to. After realizing this limitation I went back to using all of the ten countries selected to make comparisons starting with the following chart:



The above chart led to an interesting observation where the life expectancy makes a large jump at what would appear to be about six years of schooling. Using the numbers from the data set to do the averages, a more accurate number would be 7.4 years of schooling on average is when a country can expect to see around 20 years of life expectancy added to the people that live in the country. There was one country out of this set that did not follow this rule and that would be Nigeria. Before I could make any conclusions, I decided to make the chart easier to read and work with by turning it into:



Showing the change over time made the information clearer to work with and adding an overlay to make the chart interactive with hover data made tracking my findings easier. With the outlier of Nigeria in mind I was still confident in my finding that schooling does have a positive impact on life expectancy, but there are other factors that have a greater impact negatively that must be present in countries like Nigeria that are not present in countries like the UAE.

IV. Conclusion

A. Call to Action

The data collected and analyzed during this project can act as a stepping stone to further research, which can in turn be used to help educate the public about making healthier and more enriching life choices. No one variable or choice of life can determine a specific outcome, so it is important that the public be aware of how various factors come together to make a better tomorrow for themselves and for their society.

B. Bias and Limitations

The data only covers a set amount of time, from the years 2002-2019, so not only is it not entirely current, it could potentially be outdated. The data also is limited by its geographical coverage. It did have 79 countries, but not only is this not every country, the countries included could have limited data. The data also lacks a separation into desirable categories such as sex, race, ethnicity, religion, age, employment status, or political violence.

C. Future work

In the future, with more time and greater resources, datasets on this topic could be merged with new information collected regarding the variables mentioned already: sex, race, ethnicity, etc. This new data could be used to assist governments and public health officials in predicting, addressing, and preventing adverse health behaviors.

Works Cited

"*About the UAE Economy*." UAE Official Website, u.ae, https://u.ae/en/about-the-uae/economy.

Bezy, Judith Marie. "life expectancy". *Encyclopedia Britannica*, 31 May. 2024, https://www.britannica.com/science/life-expectancy. Accessed 9 June 2024.

Bowman, Phil. "Life Expectancy: Exploratory Data Analysis." Kaggle, 2022, https://www.kaggle.com/code/philbowman212/life-expectancy-exploratory-data-analysis/notebook. Accessed 9 June 2024.

Bu, Te et al. “Relationship Between National Economic Development and Body Mass Index in Chinese Children and Adolescents Aged 5-19 From 1986 to 2019.” *Frontiers in pediatrics* vol. 9 671504. 27 Apr. 2021, doi:10.3389/fped.2021.671504

Central Intelligence Agency. *The World Factbook 2021*. Washington, DC: Central Intelligence Agency, 2021. https://www.cia.gov/.

Fox, Ashley et al. “What is driving global obesity trends? Globalization or "modernization"?.” *Globalization and health* vol. 15,1 32. 27 Apr. 2019, doi:10.1186/s12992-019-0457-y

“GDP and Spending - Gross Domestic Product (GDP) - OECD Data.” *theOECD*, data.oecd.org/gdp/gross-domestic-product-gdp.htm#:~:text=Gross%20domestic%20product%20(GDP)%20is,and%20services%20(less%20imports). Accessed 17 June 2024.

Khazaei, Salman et al. “Suicide rate in relation to the Human Development Index and other health related factors: A global ecological study from 91 countries.” *Journal of epidemiology and global health* vol. 7,2 (2017): 131-134. doi:10.1016/j.jegh.2016.12.002

Meda N, Miola A, Slongo I, Zordan MA, Sambataro F. The impact of macroeconomic factors on suicide in 175 countries over 27 years. Suicide Life Threat Behav. 2022 Feb;52(1):49-58. doi: 10.1111/sltb.12773. Epub 2021 May 25. PMID: 34032310; PMCID: PMC9292781.

OpenAI. "ChatGPT." ChatGPT, OpenAI, 2024, https://chat.openai.com/. Accessed 2 June 2024 - 9 June 2024.