Analysis of Education on Life Expectancy at Birth

Introduction

In this paper, I will explore the relationship between a country's education level and its life expectancy at birth. Life expectancy is one of the key measures of population health. It is a comprehensive prediction that encompasses the mortality across a lifetime. This variable will holistically reflect a country's overall health and considers socio-economic conditions, healthcare infrastructure, cultural health practices, etc. Thus, looking at trends in a country's life expectancy at birth provides valuable insight into how that country's current living standards have improved.

Global life expectancy in general has accelerated dramatically in modern times. Revolutions in technology, medicine, education, among other factors have dramatically improved human life and longevity. Many of these factors are inter-related, and all are worth researching, although this paper will specifically look at education. Through statistical analyses, I will determine if and to what degree a country's education level has contributed to their increase in life expectancy. By doing so, we can start to make decisions on the importance of education on a global scale, and its impact it has on humanity.

Research Design

The broadest unit of analysis possible for this paper is global - life expectancy at birth (years) and education levels.

The dependent variable I am using is the global life expectancy at birth over time. The data for this is taken from the United Nations Department of Economic and Social Affairs, and they define "life expectancy at birth (years)" as: The number of years a newborn infant could expect to live if prevailing patterns of age-specific mortality rates at the time of birth stay the

same throughout the infant's life. The data is by country from the year 1990 to the year 2019. In order to analyze the global life expectancy, one could simply find the mean of life expectancy after aggregating all the countries.

The strengths of using life expectancy at birth (years) are that it considers the various factors in a country that promotes general wellbeing and health. The strengths of this particular data is that it is comprehensive; every country is represented thoroughly across the time frame with very little missing data.

The biggest weakness of using life expectancy at birth (years) as a measure of a country's health is that it is entirely predictive. A developing nation, for example, will experience a far more drastic increase in life expectancy across a lifetime than a developed nation, which ironically often have other forces that can decrease life expectancy across a lifetime, such as overconsumption. While this, and other factors, are all considered when calculating the life expectancy at birth (years), it still leaves room for doubt as to the validity of this metric.

The data was an excel spreadsheet, and was originally a csv file. Unfortunately, I had issues importing the file so I re-saved it as an excel file, and imported it that way, before turning it into a pandas data frame. The spreadsheet had to be cleaned, as the first five rows do not contain any data, rather some brief information about the data, and had to be skipped when importing. More importantly, there were holes in the data – denoted by ".." which had to turn into "NA" values. Lastly, I had to get rid of the blank columns, which were originally implemented in the excel spreadsheet for clarity, and put it into a pandas data frame. Now the education data is clean and ready to use.

The key explanatory variable I am using is the global education level over time. The data for this is taken from the United Nations. In order to "quantify" a country's education level, the United Nations based it off of the Educational, Scientific and Cultural Organization specialized agency by taking the average of mean years of schooling for adults and expected years of schooling for children, both of which are expressed as an index obtained by scaling with the corresponding maxima. In order to analyze the global life expectancy, one could simply find the mean education level after aggregating all the individual countries and dividing it by the number of countries.

The greatest strength of this variable was how thorough the data is. Education is a relatively difficult thing to quantify, although expected years of schooling is able to factor in many circumstances in a country – cultural, political, economic. Moreover, the data is inter-generational, as it not only considers how educated the adults in a country are (years of schooling they received), but it also considers the educational opportunities of the children (expected years of schooling).

Similarly to the weakness of the dependent variable, one of the two factors in determining the educational level is predictive – the expected years of schooling for children. This is quite complex to quantify, and there will always be doubts as to the accuracy of the prediction.

The data was also an excel spreadsheet, and I had to clean the data just as I did for the dependent variable.

I had two control variables – global socio-economic sustainability across a time frame, and global human security across a time frame. In order to calculate socio-economic

sustainability for a country, we use the gross capital formation of a country as percentage of their GDP. This consists of outlays on additions to the fixed assets of the economy plus net changes in the level of inventories. Fixed assets include land improvements (such as fences, ditches and drains); plant, machinery and equipment purchases; and construction of roads, railways and the like, including schools, offices, hospitals, private residential dwellings and commercial and industrial buildings. Inventories are stocks of goods held by firms to meet temporary or unexpected fluctuations in production or sales as well as goods that are work in progress. Net acquisitions of valuables are also considered capital formation. All of this data was taken from the World Bank, and the global socio-economic sustainability can be calculated by finding the mean after aggregating each country.

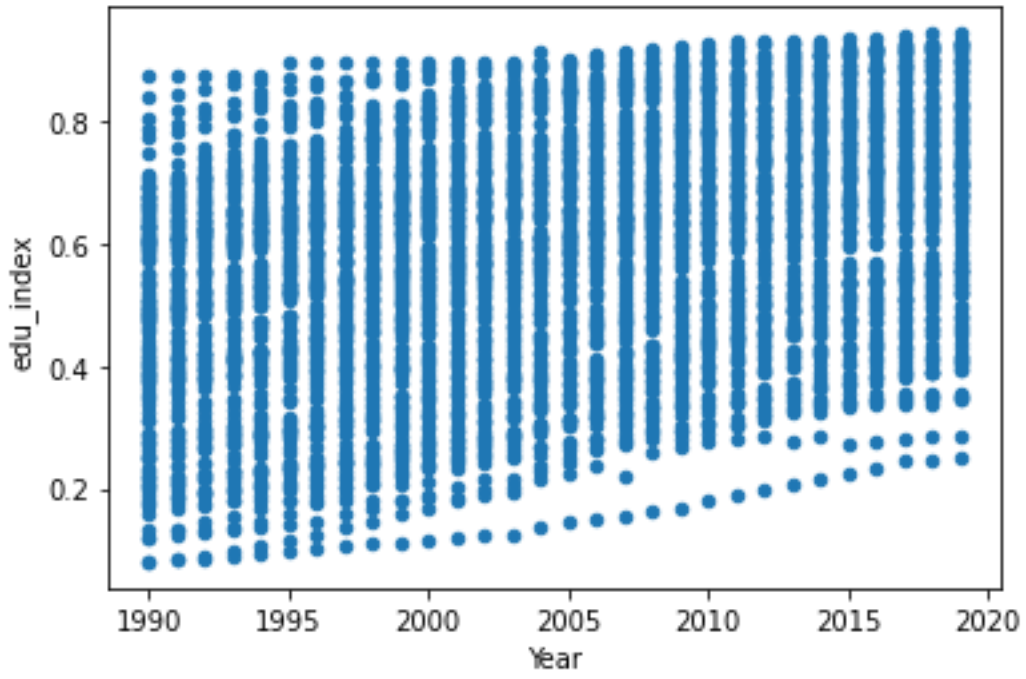
The strength of this variable is that the way it is calculated is extremely comprehensive. It clearly defines various parameters, and accurately reflects a country's general socio-economic sustainability. The only weakness is that it may be difficult to gather all of this data for each individual parameter. Thus, this may be a reason why there were more holes in the data. Not only did some countries have far less data, but some years were also completely omitted. This data was more annoying to clean, as I had to manually adjust to the missing years. However, other than that, everything else remained the same.

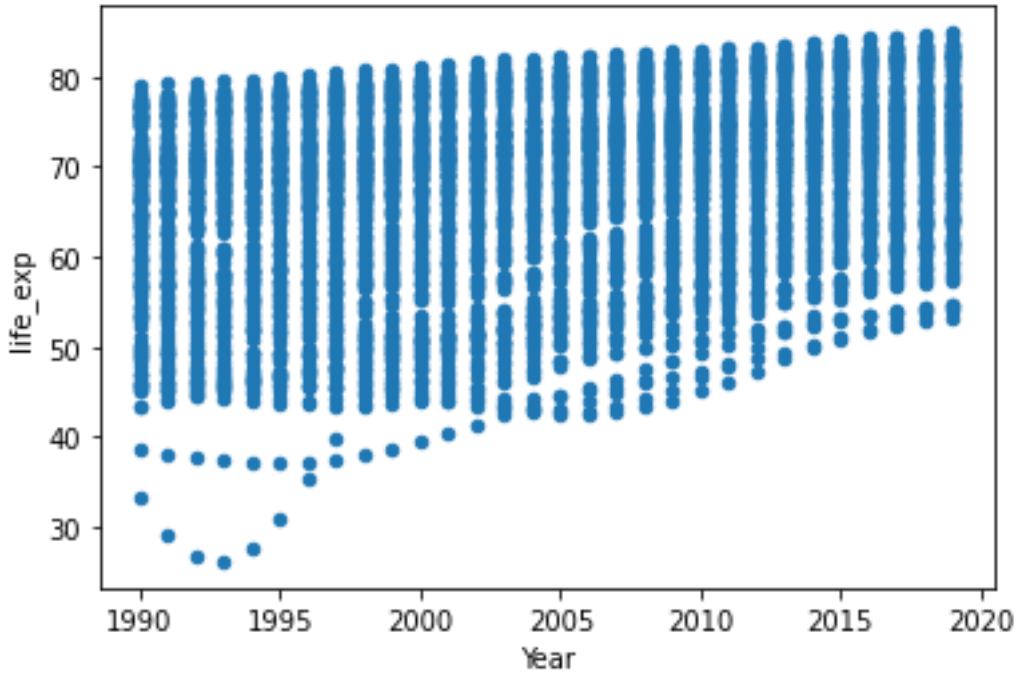
The reason it is a suitable control variable is that it also serves as a factor for life expectancy. Generally speaking, it is not just rich countries that have higher life expectancies (infrastructure, medicine/health care, cleaner cities) but rich countries that continue to grow, ie their socioeconomic sustainability. This means reinvestment into capital to continue growing the economy, and thus, contributing to increasing life expectancy.

The other control variable I used was human security. In order to calculate this, I took data from the United Nations Office on Drugs and Crime on the homicide rate (per 100,000 people). They define this as: the number of unlawful deaths inflicted upon a person with the intent to cause death or serious injury, expressed per 100,000 people. The strength of this is that it does a great job of quantifying a country's security. Given that homicide is the most extreme form of violence, this data will generally reflect the local crime environment. Violence and crime obviously decreases life expectancy, although it often seems to be hard to precisely measure the homicides in a year. This was the biggest weakness, as there were far fewer years with data. To give a reference, there was around 75% less data in for this than there was for education or life expectancy. This made coding it a hassle, and I had to think of a way around it when controlling for the variable.

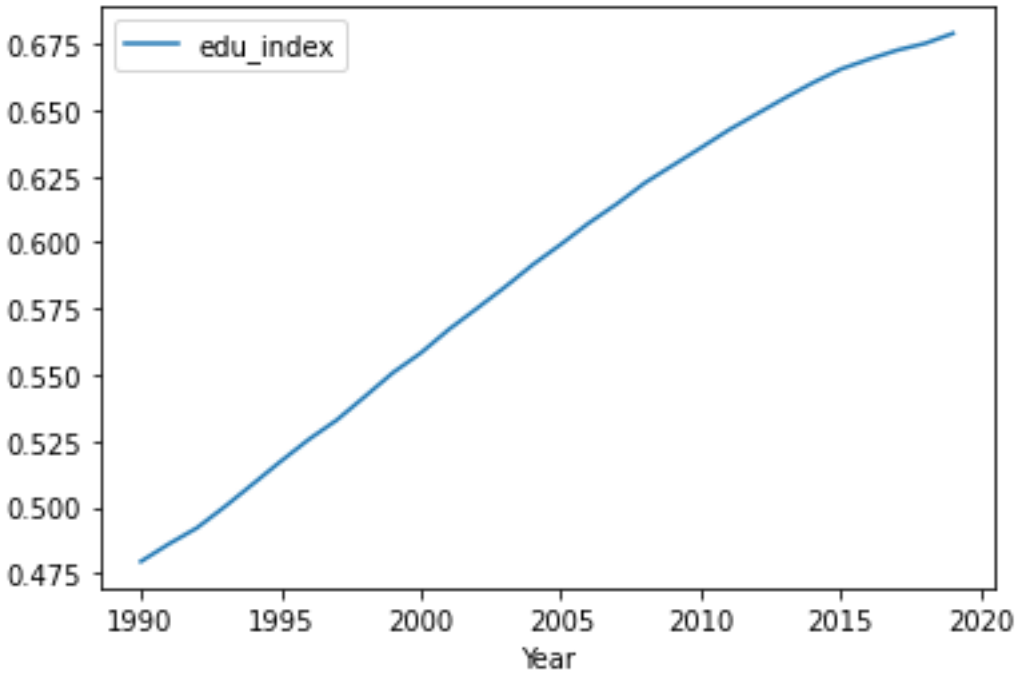
Lastly, I used various plotting methods and graphs, as well as a regression. I used an ordinary least squares regression because I had no binary response variables. Everything was a scalar, and fell within similar ranges. Also, ordinary least squares regressions are designed to model continuous response variables. For this paper, the response variable was life expectancy at birth (in years), which is certainly not a binary value. Thus, I used an ordinary least squares regression.

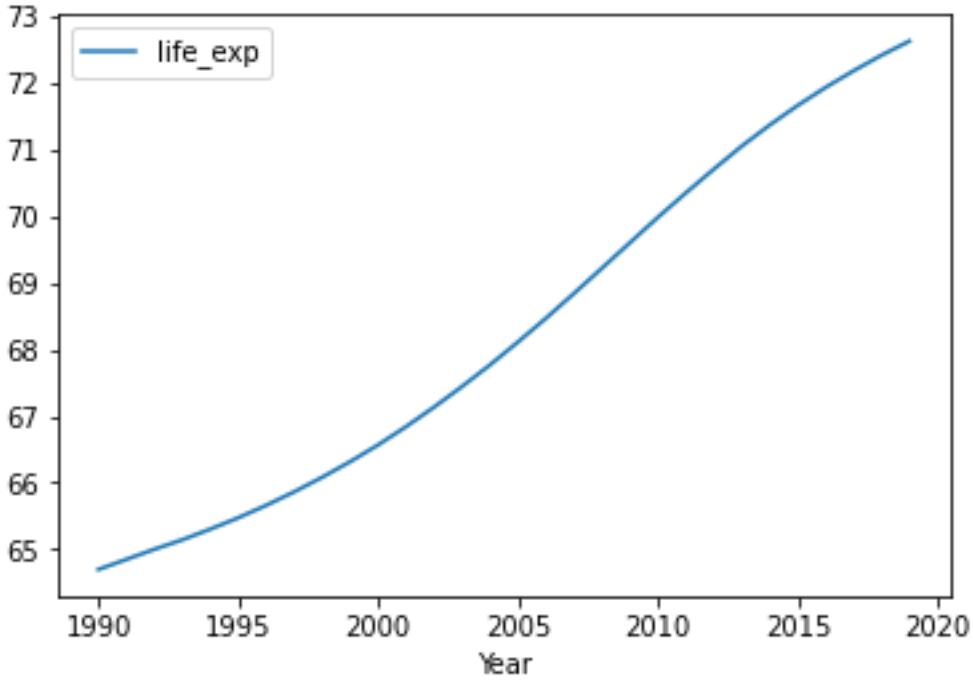
Analysis and Discussion





The first graph shows a scatterplot of the general trend of global education across time (1990-2020). As you can see, it is very hard to tell what is happening here. Similarly, the second scatterplot that shows the general trend of life expectancy at birth (in years) is hard to read.

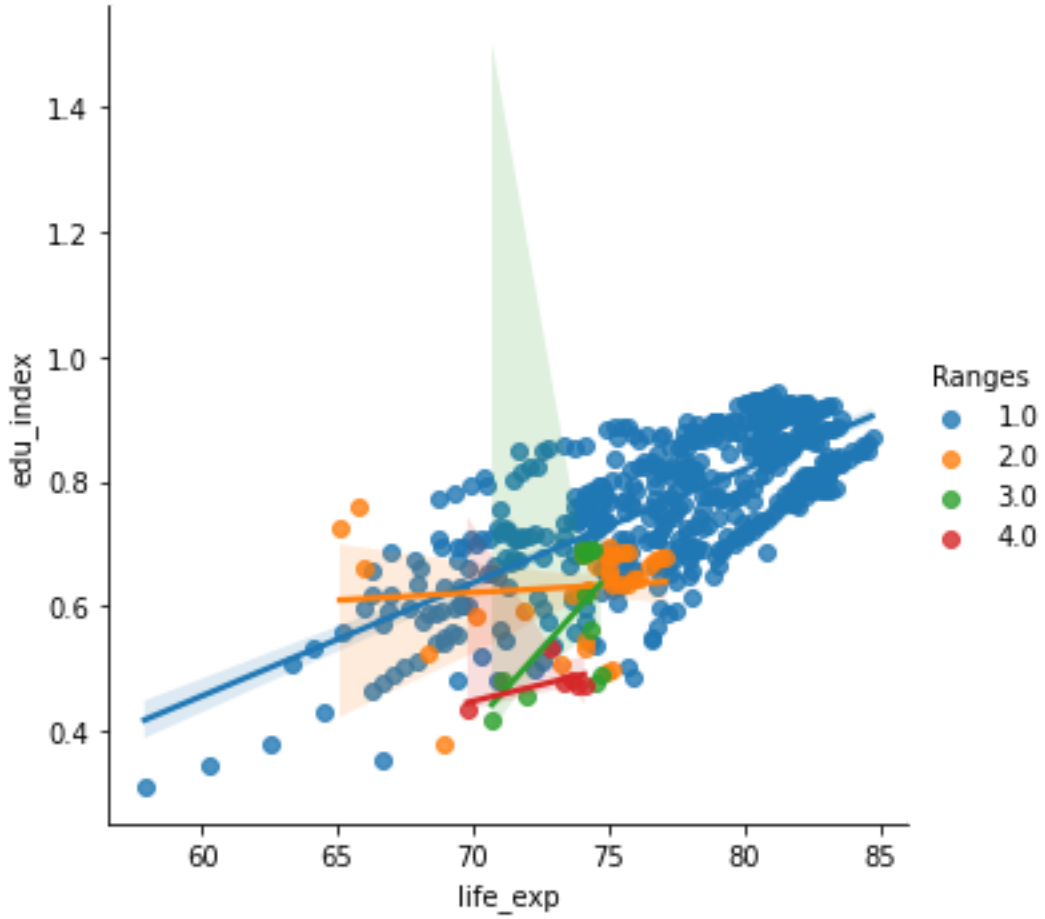


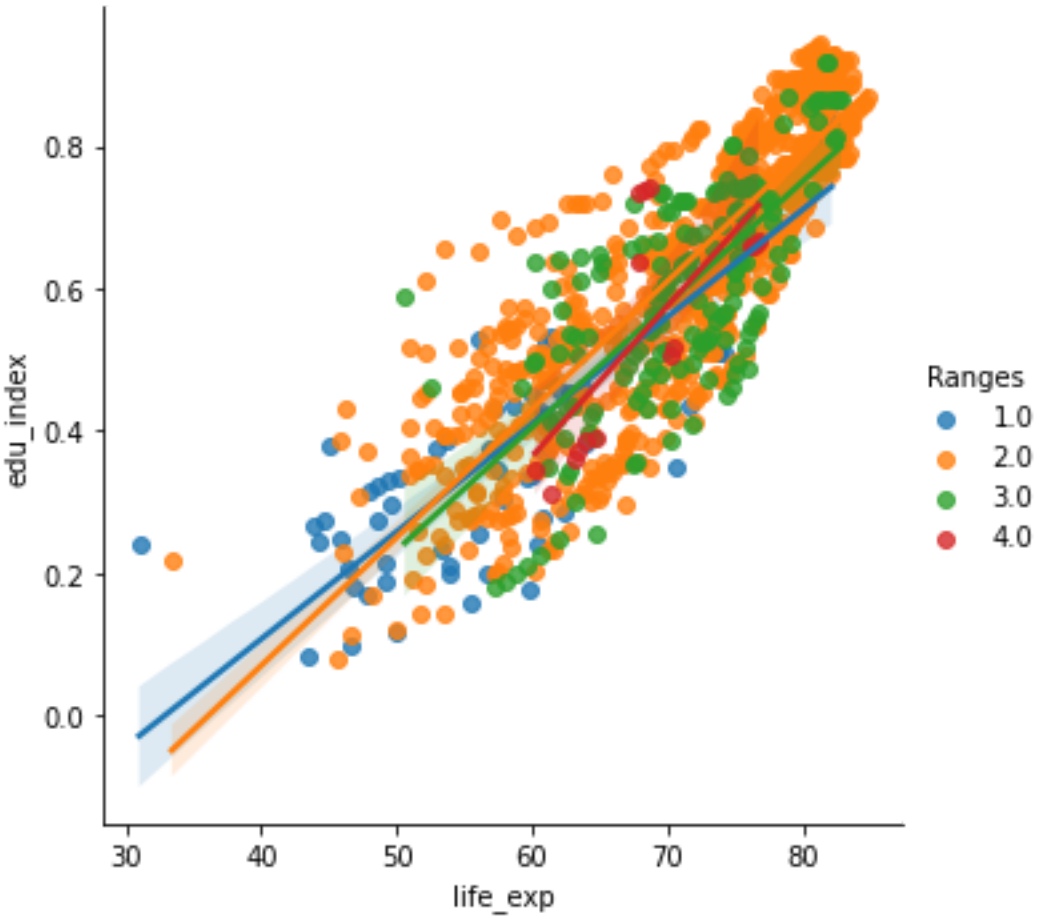


However, when I use a line graph, you can begin to see a correlation forming. The curves seem very similar, and imply a relationship between education and life expectancy.

	Coefficient	Std Err	T statistic	P-value (two-tailed)	95% CI	
					Lower	Upper
Life Expectancy	44.7597	0.285	157.263	< 0.0001	44.202	45.318
Education	41.1661	0.459	89.621	< 0.0001	40.266	42.067

This is the bivariate regression of life expectancy against education. This shows a positive relationship between life expectancy and education, as education increases, so while life expectancy. The relationship between the two variables is statistically significant because the P-value is <0.0001.





These are the two multivariable regressions. The first controls for homicide rate, the second for socioeconomic sustainability. As you can see, there are relatively high correlations with each control to life expectancy. The ranges I have used stratify the controls into four segments (quartiles). The first segments countries based on their violence (four categories), the second segments countries based on the socio-economic sustainability. Both ranges were calculated in comparison to the most violent country and the most socio-economically sustainable, respectively.

Thus from both all of the graphs and data, we see a clear relationship between education and health. However, at the same time, the controls greatly influence the dependent variable. Overall, it is safe to say that global education level strongly influences the global life expectancy.

Conclusion

In this study, I examined the relationship between education and health. I used education levels as calculated by the United Nations, and life expectancy at birth (in years).

My conclusion is about what I expected. Of course, there would be a strong relationship between my chosen explanatory variable and life expectancy, and of course there are many factors (like my controls) that play a big role. Nonetheless, the takeaways should be that investing in education seems like a laudable goal to increase global life expectancy.

Technical Appendix

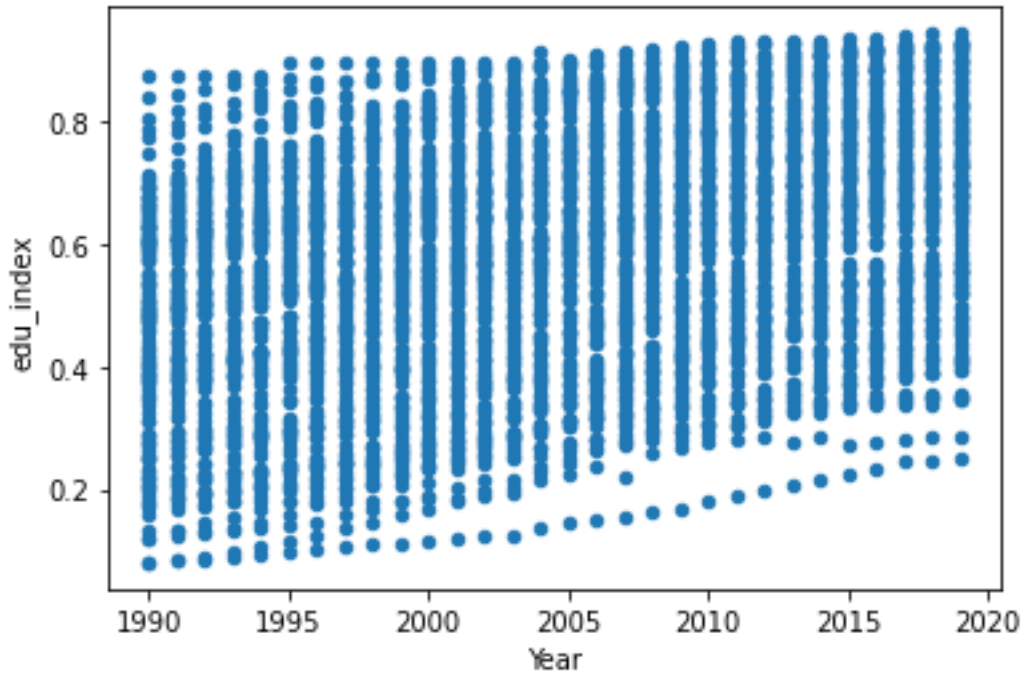
```
# %%
import pandas as pd
import numpy as np
import seaborn as sns
import statsmodels.api as sm
import statsmodels.formula.api as smf
import matplotlib.pyplot as plt
import math

# %%
from google.colab import files
uploaded = files.upload()

# %%
ed_data = pd.read_excel("Education index.xlsx", skiprows = 5, na_values = '..')
life_data = pd.read_excel("Life expectancy at birth (years).xlsx", skiprows = 6,
na_values = '..')
homicide_data = pd.read_excel("Homicide rate (per 100,000 people).xlsx", skiprows
= 6, na_values = '..')
gdp_data = pd.read_excel("Gross capital formation (% of GDP).xlsx", skiprows = 5,
na_values = '..')

# %%
ed_data = ed_data.iloc[:,2:62:2]
l = pd.read_excel("Education index.xlsx", skiprows = 5, usecols = [0,1])
l
ed_data = l.join(ed_data)
ed_data = ed_data.dropna()
ed_data = ed_data.reset_index()

# %%
ed_data.tail

# %%
ed_data = ed_data.dropna()

# %%
ed_data.info

# %%


# %%


# %%
```

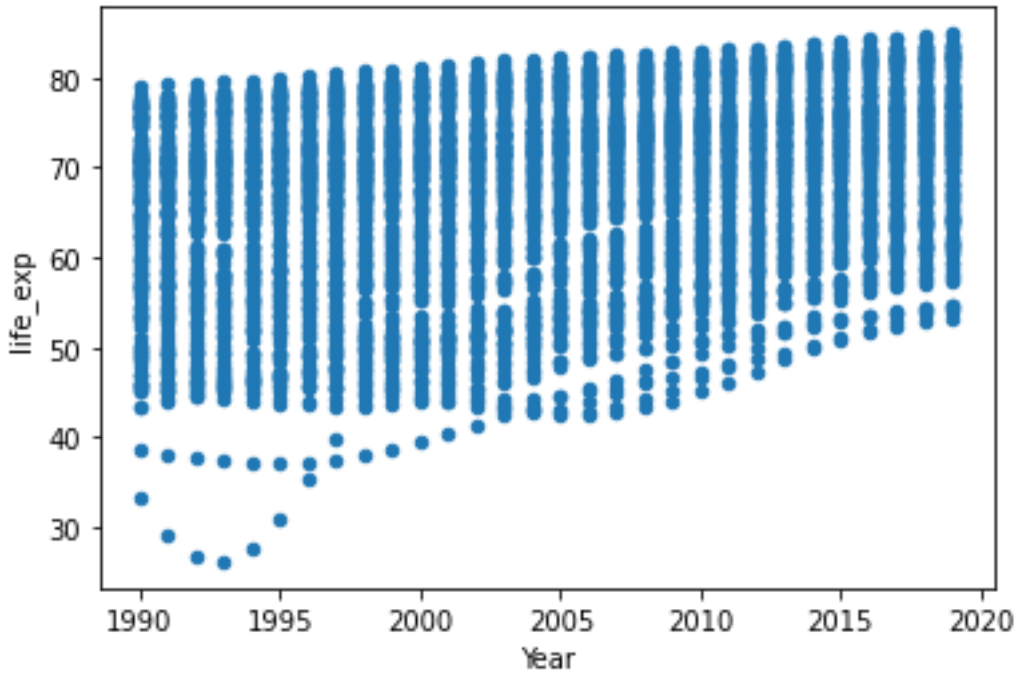
```
life_data = life_data.iloc[:,2:62:2]
l = pd.read_excel("Life expectancy at birth (years).xlsx", skiprows = 6, usecols
= [0,1])

life_data = l.join(life_data)
life_data = life_data.dropna()
life_data = life_data.reset_index()

# %%
homicide_data = homicide_data.iloc[:,2:30:2]
l = pd.read_excel("Homicide rate (per 100,000 people).xlsx", skiprows = 6,
usecols = [0,1])
l
homicide_data = l.join(homicide_data)
homicide_data = homicide_data.dropna()
homicide_data = homicide_data.reset_index()

# %%
gdp_data = gdp_data.iloc[:,2:62:2]
l = pd.read_excel("Gross capital formation (% of GDP).xlsx", skiprows = 5,
usecols = [0,1])
l
gdp_data = l.join(gdp_data)
gdp_data = gdp_data.dropna()
gdp_data = gdp_data.reset_index()

# %%
ed_data = ed_data.append(ed_data.mean(),ignore_index = True)

# %%
ed_data.loc[144,"Country"]  = "Mean"
ed_data.tail()

# %%
ed_data =pd.melt(ed_data,id_vars = ["index","HDI Rank","Country"],value_vars
=list(range(1990,2020)),var_name = "Year",value_name = "edu_index")

# %%
ed_data.plot.scatter(x="Year",y="edu_index")

# %%
ed_data

# %%
ed_data_mean = ed_data.groupby("Year").mean().reset_index()
ed_data_mean.plot(x="Year",y="edu_index")

# %%
```

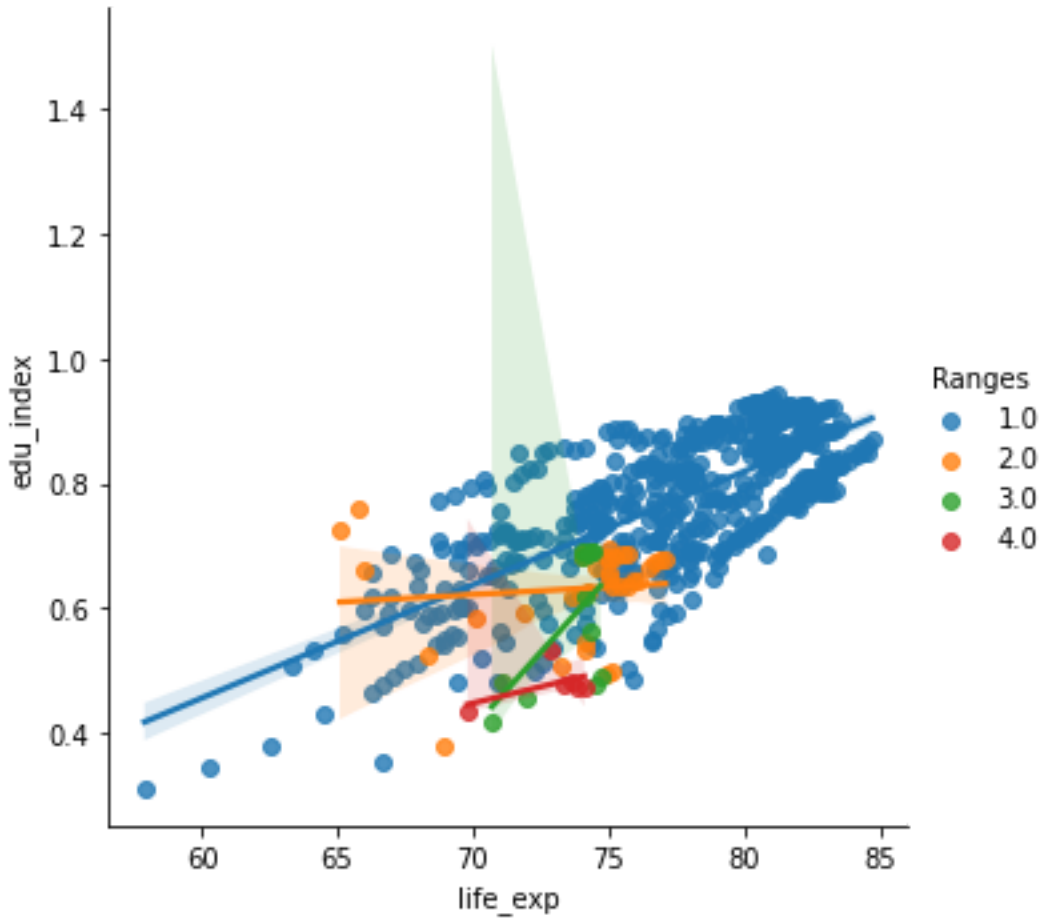
```
# %%
life_data = life_data.append(life_data.mean(),ignore_index = True)

# %%
life_data.loc[191,"Country"]  = "Mean"
life_data.tail()

# %%
life_data =pd.melt(life_data,id_vars = ["index","HDI Rank","Country"],value_vars
=list(range(1990,2020)),var_name = "Year",value_name = "life_exp")

# %%
life_data.plot.scatter(x="Year",y="life_exp")

# %%
life_data

# %%
life_data_mean = life_data.groupby("Year").mean().reset_index()
life_data_mean.plot(x="Year",y="life_exp")

# %%
homicide_data = homicide_data.append(homicide_data.mean(),ignore_index = True)

# %%
homicide_data.loc[49,"Country"]  = "Mean"
homicide_data.tail()

# %%
homicide_data = pd.melt(homicide_data,id_vars = ["index","HDI
Rank","Country"],value_vars = [1990, 1995, 2000, 2005, 2010, 2011, 2012, 2013,
2013, 2015, 2016, 2017, 2018],var_name = "Year",value_name = "hom_rate")

# %%
homicide_data.plot.scatter(x="Year",y="hom_rate")

# %%
homicide_data
homicide_data

# %%
type(homicide_data['hom_rate'])
list1 = [i * 4/83.8 for i in homicide_data['hom_rate'].tolist()]
list1
homicide_data["Ranges"] = np.ceil(list1)

# %%
```

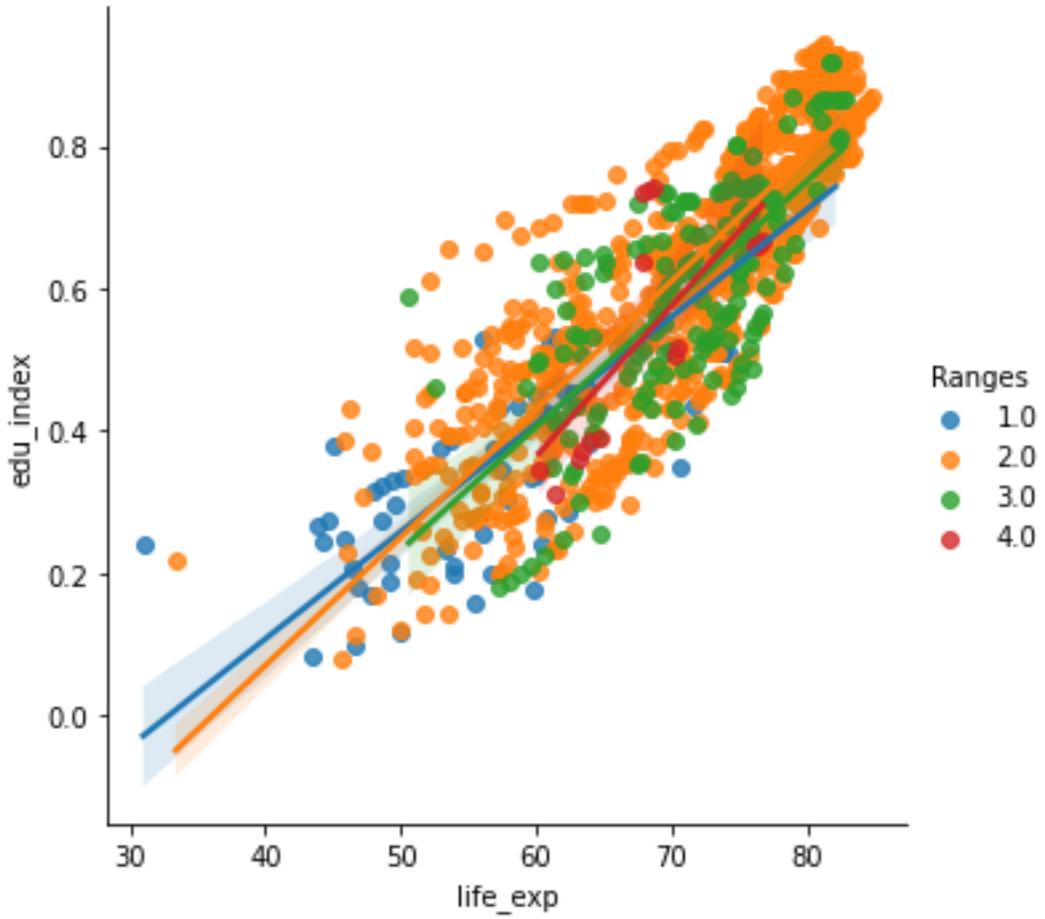
```
homicide_data["Ranges"].unique()

# %%
homicide_data_mean = homicide_data.groupby("Year").mean().reset_index()
homicide_data_mean.plot(x="Year",y="hom_rate")

# %%
list1 = [i * 4/83.8 for i in homicide_data['hom_rate'].tolist()]
list1
homicide_data["Ranges"] = np.ceil(list1)

# %%
list2 = [i * 4/58.2 for i in gdp_data['gdp'].tolist()]
list2
gdp_data["Ranges"] = np.ceil(list2)

# %%
all_data = pd.merge(ed_data, life_data, how = "left", on = ["Country", "Year"])
all_data = all_data.dropna()

# %%
type(all_data)

# %%
y = all_data['life_exp']
x = all_data['edu_index']
x = sm.add_constant(x)
model = sm.OLS(y, x).fit()
print(model.summary())

# %%
hom_control = pd.merge(homicide_data, all_data, how = "left", on = ["Country",
"Year"])
hom_control = hom_control.dropna()

# %%
sns.lmplot(data=hom_control,x="life_exp",y="edu_index",hue="Ranges")

# %%


# %%
gdp_data = gdp_data.append(gdp_data.mean(),ignore_index = True)

# %%
gdp_data.loc[104,"Country"]  = "Mean"
gdp_data.tail()
gdp_data
```

```
# %%
gdp_data
#gdp_data = pd.melt(gdp_data,id_vars = ["index","HDI Rank","Country"], value_vars
= [1990, 1995, 2000, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014,
2015, 2016, 2017, 2018, 2019],var_name = "Year",value_name = "gdp")

# %%
gdp_data.plot(x="Year",y="gdp")

# %%
gdp_data

# %%
gdp_control = pd.merge(gdp_data, all_data, how = "left", on = ["Country",
"Year"])
gdp_control = gdp_control.dropna()

# %%
sns.lmplot(data=gdp_control,x="life_exp",y="edu_index",hue="Ranges")

# %%
```




life_exp
80
70
60
50
40
30
1990
1995
2000
2005
2010
2015
2020
Year


edu_index
1.4
1.2
1.0
0.8
0.6
0.4
60
65
70
75
80
85
life_exp
Ranges
1.0
2.0
3.0
4.0


edu_index
0.0
0.2
0.4
0.6
0.8
30
40
50
60
70
80
life_exp
Ranges
1.0
2.0
3.0
4.0