

Project Edu+?

A3: Exploratory Data Analysis (02/05/2020)

HCDE 511 Information Visualization

Winter 2020

Amitabh Nag(amnag@uw.edu) | Susan Yishan Zheng (susanz96@uw.edu)

Introduction

Knowledge is power and education is key for people to improve social mobility. We are very interested in examining the historical data in world education attainment and enrollment. We also think other factors like life expectancy and suicide rate may relate to education attainment and enrollment. In addition, we want to explore how those factors affect education attainment across different regions and countries.

Dataset

Our original dataset on education attainment comes from [Barro-Lee Website](#).

In order to find factors that might relate to education attainment, we found suicide rate data and life expectancy from Our World In Data. We joined all the data tables together into one using Python and did some data cleanup using Tableau Prep.

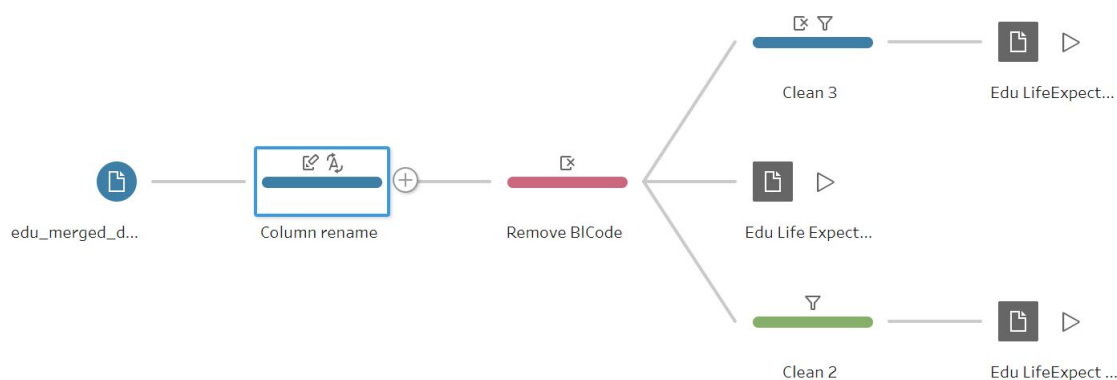
Description of data variables:

No.	Variables	Data Type	Description	% Complete
1	Human Capital	Quantitative	Human Capital in the World, age between 15-64	74%
2	Human Capital Adjusted	Quantitative	Alternative Human Capital, age between 15-64	74%
3	Total Schooling	Quantitative	In Number of Years	74%
4	No Schooling	Quantitative		74%
5	Population	Quantitative	In Thousands	74%
6	Primary Enrollment	Quantitative	Ratio	64%
7	Primary Schooling	Quantitative	Percentage	74%
8	Primary Schooling Complete	Quantitative	Percentage	74%

9	Secondary Enrollment	Quantitative	Ratio	29%
10	Secondary Schooling	Quantitative	Percentage	74%
11	Secondary Schooling Complete	Quantitative	Percentage	74%
12	Tertiary Enrollment	Quantitative	Ratio	14%
13	Tertiary Schooling	Quantitative	Percentage	74%
14	Tertiary Schooling Complete	Quantitative	Percentage	74%
15	Total Primary Schooling	Quantitative	In Number of Years	74%
16	Total Secondary Schooling	Quantitative	In Number of Years	74%
17	Total Tertiary Schooling	Quantitative	In Number of Years	74%
18	Suicide Rate	Quantitative	Percentage	13%
19	Life Expectancy	Quantitative	Percentage	44%
20	Country	Nominal	Countries of the world	100%
21	Region	Nominal	6 regions: Advanced Economies, Asia and the	100%

			Pacific, Eastern Europe, Latin America and the Caribbean, Middle East and North Africa, and Sub-Saharan Africa.	
22	Sex	Nominal	F, M, and MF	100%
23	Year	Ordinal	From 1820 to 2010	100%

Here is the flow that we used to clean the data in Tableau prep:



Initial Research Questions

1. Does education attainment have had an upward curve over the years?
2. With the increase in education attainment, does suicide rate increase for developed countries and decrease for developing countries, and does life expectancy increase for all countries?
3. Is it true that at a global level, the average life expectancy has increased for countries with higher rates of change in education attainment over a certain time period and the increase in the average life expectancies of these countries is larger during the same time period?

4. Asia-Pacific (excluding Australia and New Zealand), Africa and South America have higher rates of changes in education attainment compared to North America and Europe

Visual Exploration Process

Visual Exploration of Initial Question 1:

Question 1: Does education attainment have had an upward curve over the years?

Since this question is very broad and it can be divided into 3 sub-questions. Below are the refined questions:

1.a Does education attainment have had an upward curve overtime globally?

1.b Does education attainment have had an upward curve overtime regionally?

1.c Does education attainment have had an upward curve overtime nationally?

Iteration 1 for Question 1.a:

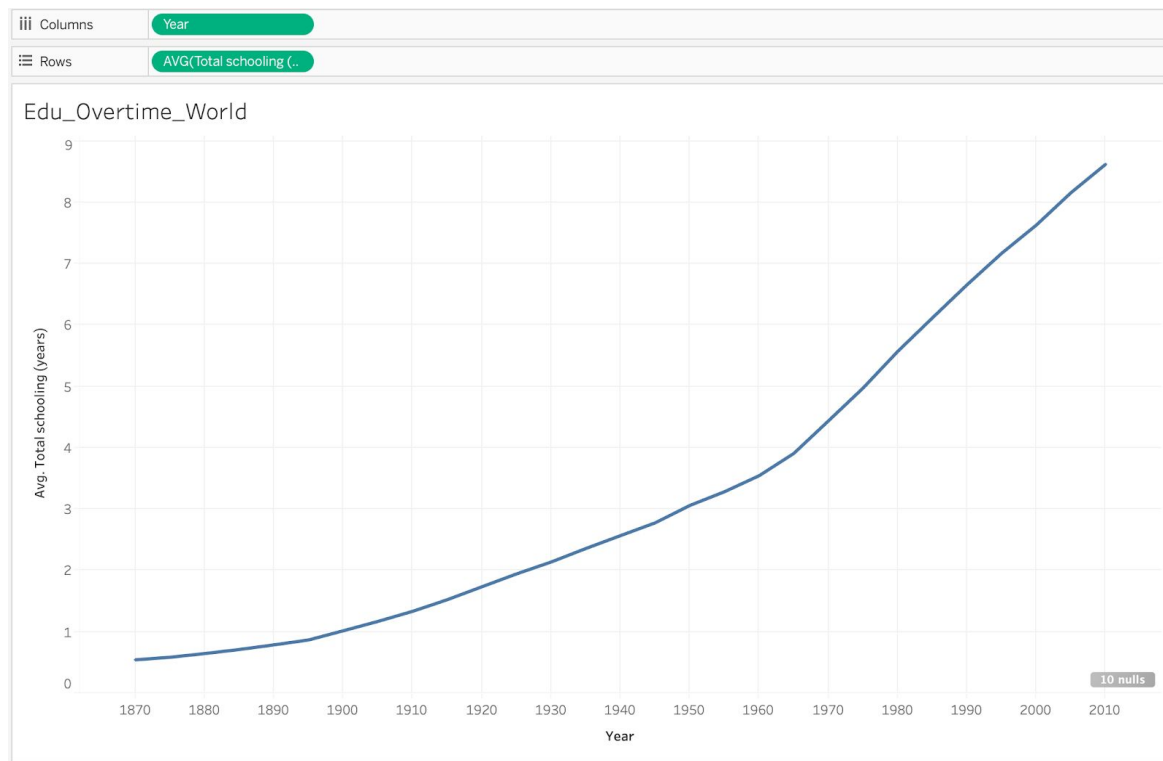


Chart type: Line Graph

Variables (encoding):

- World Average Total Schooling in Years (Y axis)
- Time (X axis)

Pros: Simple to understand the trend of world education attainment overtime

Cons: Does not tell a holistic story. Using the average values might overlook outliers

Discussion:

- Overall, this graph does give an answer to the question: education attainment has an upward curve over the years.
- However, it does not tell us much about whether the trend has remained the same across regions (continents).
- Need to visualize the trends of total schooling in years for all regions.

Iteration 2 for Question 1.b:

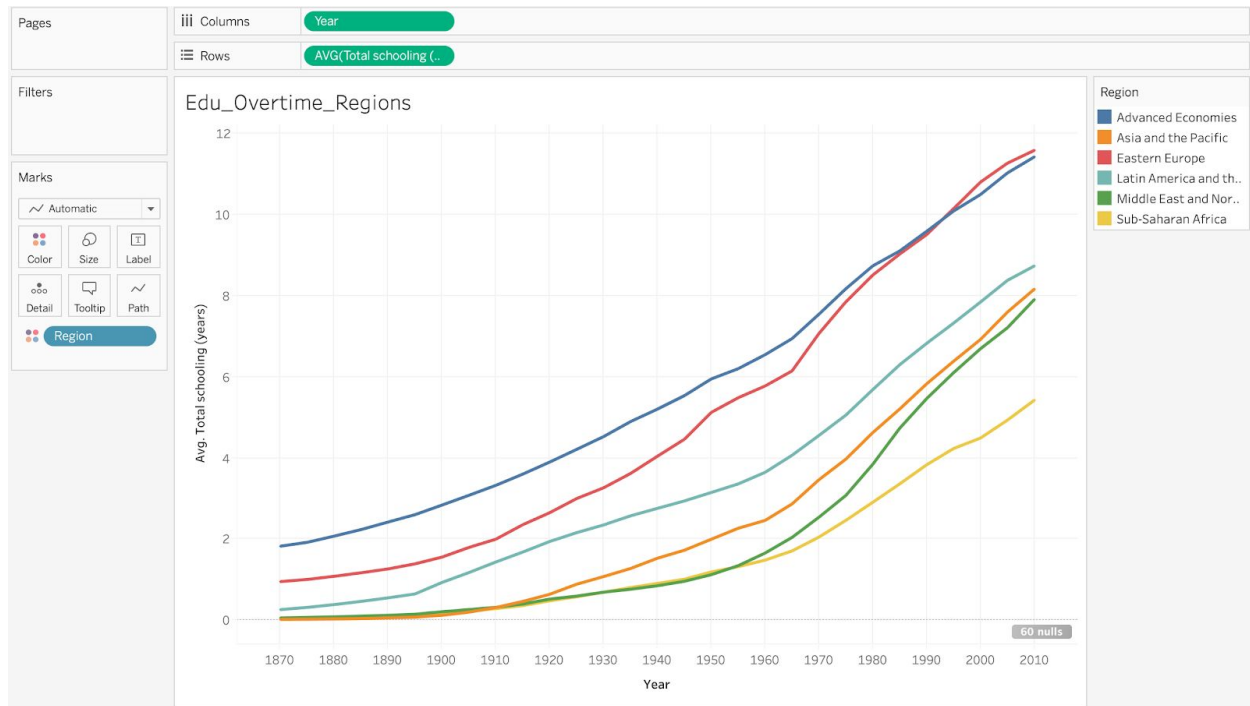


Chart type: Line Graph

Variables (encoding):

- Average Total Schooling in Years (Y axis)
- Time (X axis)
- Regions (Advanced Economies, Asia and the Pacific, Eastern Europe, Latin American and the Caribbean, Middle East and North Africa, and Sub-Saharan Africa)

Pros: Simple to understand the trend of education attainment overtime across regions

Cons: Not able to tell what specific countries had been driving the growth of total schooling within the regions overtime.

Discussion:

- Overall, this graph does give an answer to the question: education attainment has an upward curve over the years among regions.
- This visualization makes sense to us and it generally matches with our expectation; however we still need to uncover more details:
 - Advanced Economies have had the highest total schooling in years on average overtime (till 1990); after 1990, total schooling on average for Eastern Europe exceeded Advanced Economies – more exploration needed to show which specific countries in Eastern Europe drove the growth of total schooling.

- Before 1955, the Middle East & North Africa, and Sub-Saharan Africa had similar trends – their lines are overlapping. After 1955, the line for Middle East & North Africa went upward exponentially. It will be interesting to explore what had driven the growth of total schooling in that region.
- Need to visualize the trends of total schooling in years for selected countries.
 - For next iteration, try to reveal what countries in Eastern Europe, and in the Middle East & North Africa region had the highest total schooling in years on average.
 - After the next iteration, try to pick several countries within a region to compare their trends. (Especially from Eastern Europe, and Middle East & North Africa)

Iteration 3:

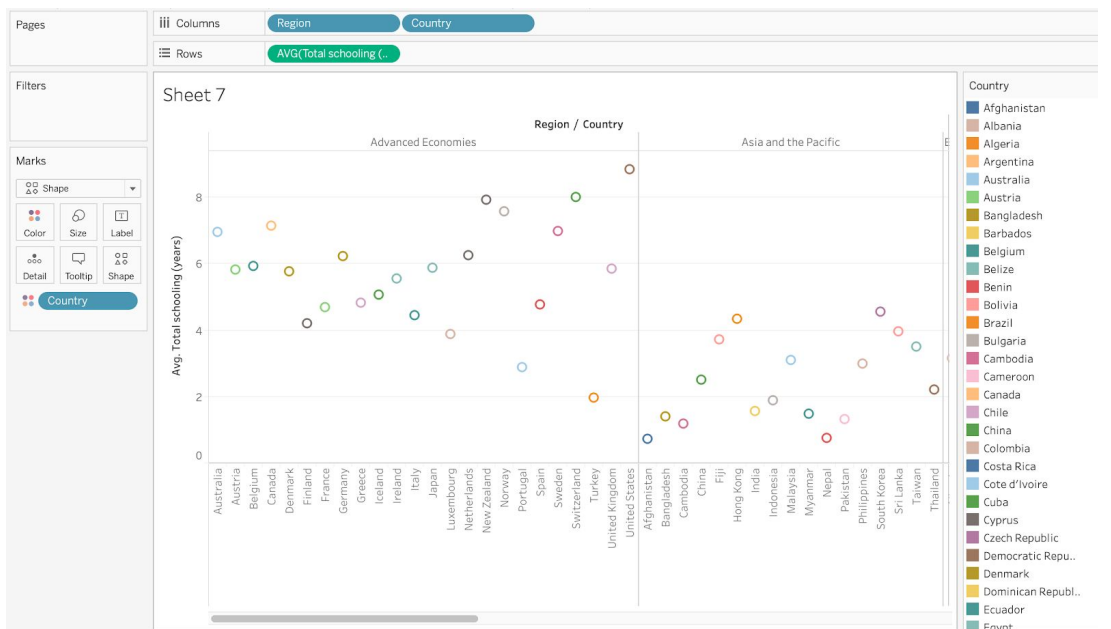




Chart Type: Scatterplot, compared side-by-side

Variables (encoding):

- Regions & Countries (horizontal)
- Average Total Schooling in Years

Pros: shows me which countries/sub regions within a region have the highest average total schooling, which helps further visual exploration.

Cons: N/A

Discussion:

- In Eastern Europe, Czech Republic has the highest average total schooling(7.211 years), Hungary has the second highest(6.334 years), and Poland has the third highest(5.408 years). The lowest is that of Albania(3.159 years). – hover on Tableau to see years
- In the Advanced Economies region, the United States has the highest average total schooling(8.828 years), the second is Switzerland (8.008 years), and the third is New Zealand(7.934). The lowest is Turkey, which has an average of 1.98 years of total schooling.
- In Asia and the Pacific region, the highest is South Korea (4.556 years), the second is Hong Kong, China (4.343 years), Sri Lanka is the third highest (3.957 years). China has an average of 2.497 years in total schooling and Indian has an average of 1.56 years. The lowest is Afghanistan (0.734 years).

- In Latin America and the Caribbean region, Belize is the highest (5.336 years), Trinidad and Tobago has the second highest (4.787 years), and Barbados has the third highest(4.71 years). The lowest is Haiti, which has an average of 1.29 years.
- In the Middle East and North Africa region, Cyprus has the highest(4.349 years), Malta has the second highest(4.167 years), and Yemen has the lowest(0.473 years).
- In the Sub-Saharan Africa region, the highest is South Africa (3.794 years), Mauritius has the second highest (3.244 years), and Mali has the lowest (0.375 years).
- For next iteration, I want to visualize the trends in these selected countries/region:
 - Czech Republic
 - Hungary
 - Poland
 - The US
 - Switzerland
 - New Zealand
 - South Korea
 - Hong Kong, China and Greater China
 - India
 - Afghanistan
 - Belize
 - Trinidad and Tobago
 - Cyprus
 - Yemen
 - South Africa
 - Mali

Iteration 4 for Question 1.c:

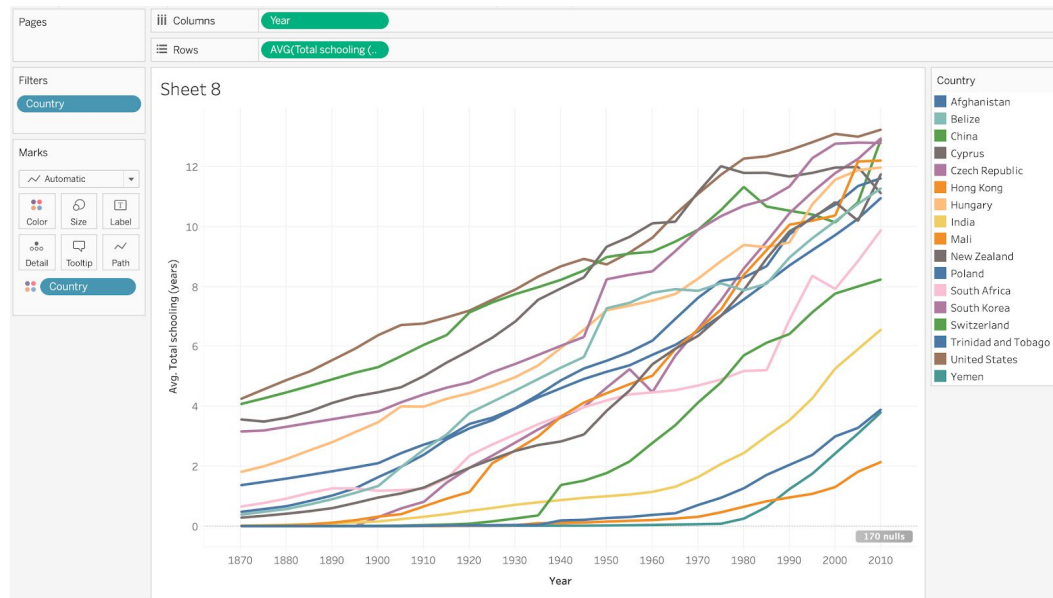


Chart Type: Line Graph

Variables (encoding):

- Year (X Axis)
- Average Total Schooling in Years (Education Attainment) (Y Axis)
- Countries/special zones

Pros: Reveals the need to separate the lines for different regions

Cons: The lines are too cluttered, and it's hard to distinguish among the lines in the upper right corner.

Discussion:

- From this visualization, we can at least infer that Czech Republic had driven the growth of total schooling in Eastern Europe overtime.
- South Africa had a rapid growth in total schooling from 1985 to 1990, and its line remained above China's after 1990.
- Advanced economies such as the US, New Zealand, and Switzerland overall have upward line trends; however Switzerland experienced a sudden drop in total schooling starting 1980, and it kept declining until 2000. Average total schooling in New Zealand had remained almost the same from 1975 to 2005, and it dropped after 2005.
- Need to separate the lines for countries based on different regions.

Iteration 5 for Question 1.b and 1.c:

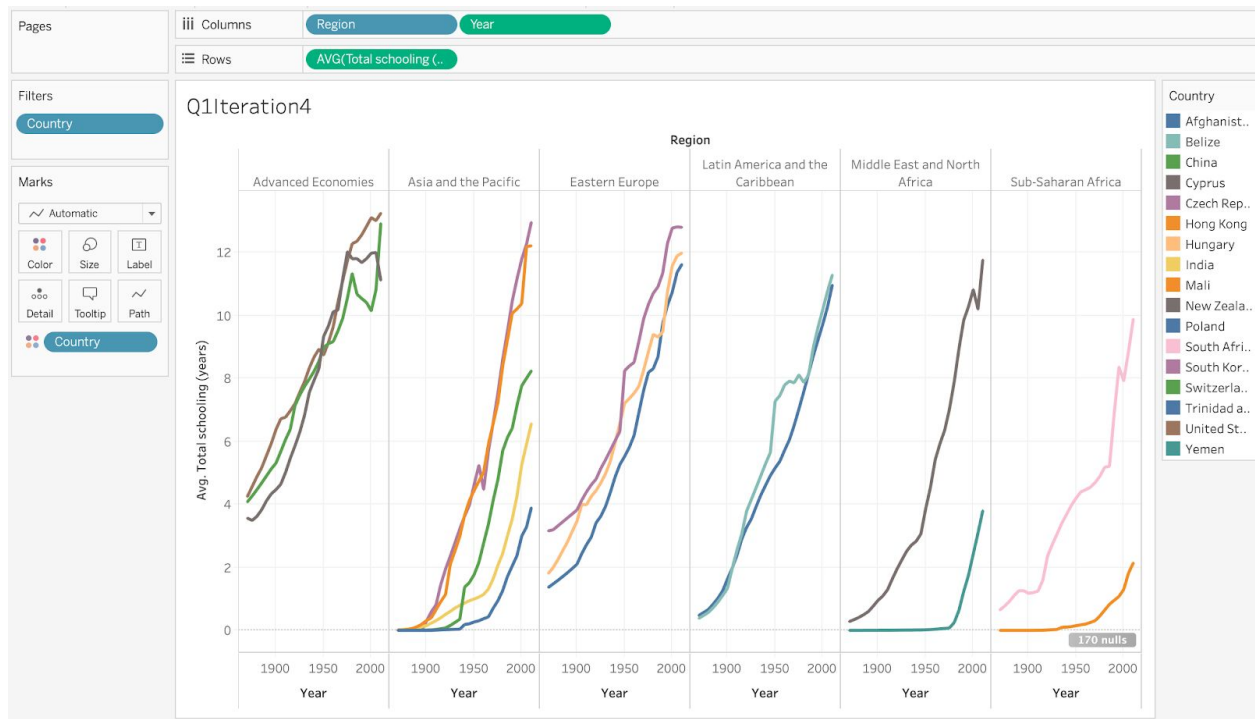


Chart Type: Segmented Line Graph

Variables (encoding):

- Year – Horizontal
- Region – Side-by-side Comparison
- Average Schooling in Years

Pros: provides a clear comparison of total schooling in years of selected countries across different regions.

Cons: Each region has a limited width on the graph, and it's hard to compare the lines in a region. It would be nicer if there's a scroll bar under the graph and we can adjust the width of each region displayed.

Discussion:

- Apparently, the United States in Advanced Economies region have the highest total schooling in years; both New Zealand and Switzerland experienced a decline in total schooling around the same time (in 1975-1980). We can investigate what might cause those decline in those two countries, or explore the factors that influenced the total schooling in years around that time.
- In the Asia and the Pacific region, most countries have an upward curve overtime; For South Korea, its total schooling in years declined from 1955 to 1960. We can also look into the factors that influenced the decline in total years of schooling.

- In the Eastern Europe region, we can attest that Czech Republic drove the growth of total schooling within the region, which might be the reason why the average total schooling in years of Eastern Europe exceeded that of Advanced Economies after 1990.

Question 2

With the increase in education attainment, does suicide rate increases for developed countries and decreases for developing countries, and does life expectancy increases for all countries?

This question contains three sub-questions, hence splitting this into three parts:

2.a With the increase in education attainment, does suicide rate increases for developed countries?

2.b With the increase in education attainment, does suicide rate increases for developing countries?

2.c With the increase in education attainment, does life expectancy increase for all countries?

Iteration 1 for Questions 2.a and 2.b Combined:

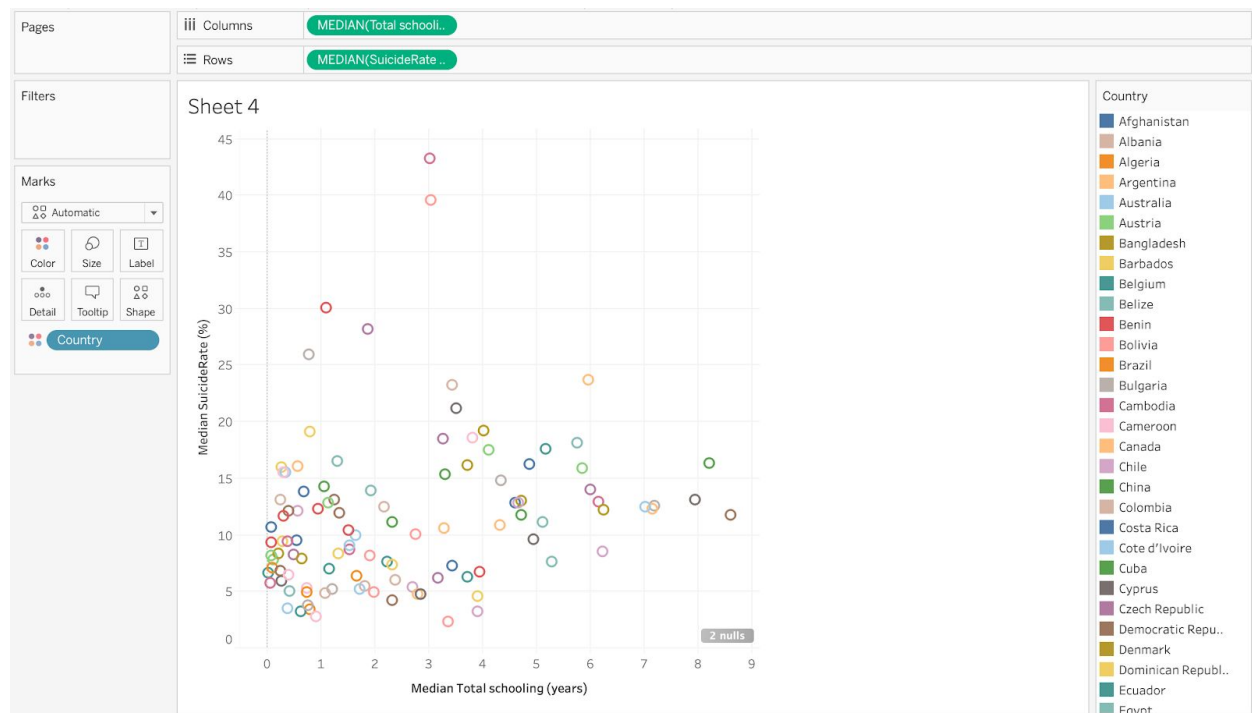


Chart Type: Scatter Plot

Variables (encoding):

- Median Total Schooling in Years (X Axis)
- Median Suicide Rate (Y Axis)
- Mark: countries in different colors

Pros: we found out since we can't fit a trend line in the scatterplot, we need to separate the countries into groups.

Cons: Can't tell if there's a correlation between median suicide rate and total schooling in years.

Discussion:

- More visual exploration is needed to examine the relationship between suicide rate and total schooling in years (education attainment)

Iteration 2 for Question 2.a:

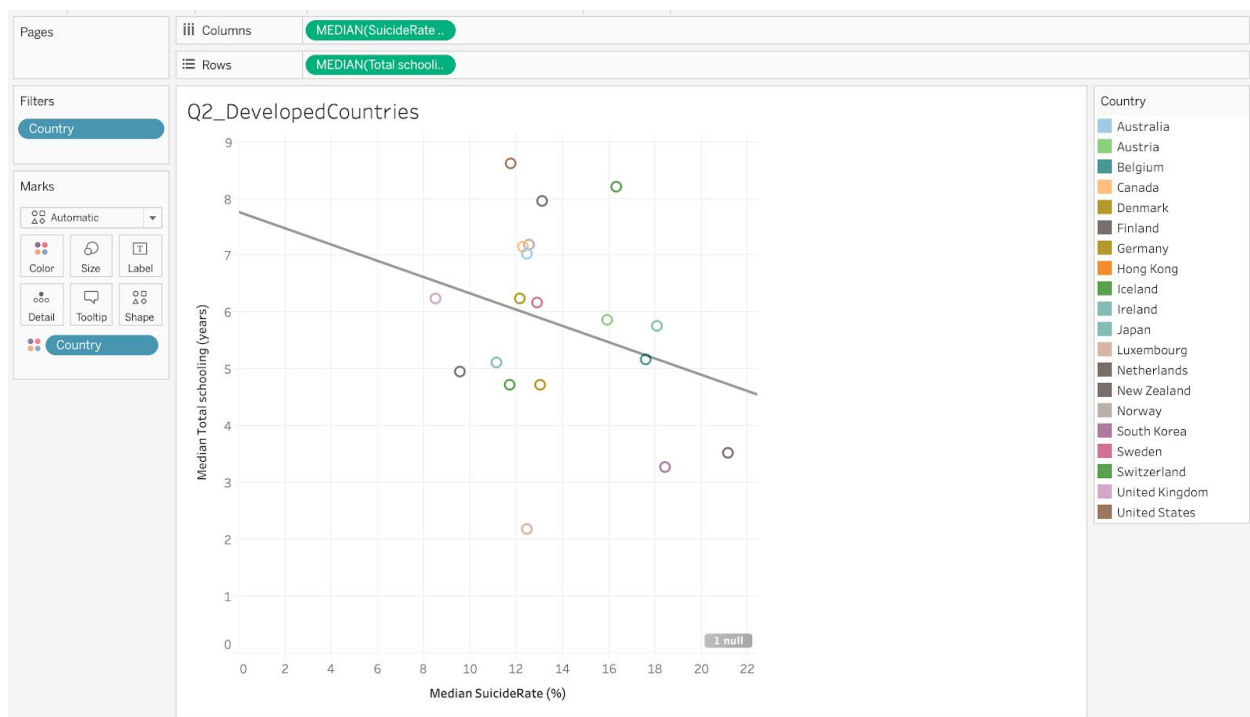


Chart Type: Scatterplot

Variables encoding:

- Median Suicide Rate (X Axis)
- Median Total Schooling in Years (Y Axis)
- Mark: Selected Developed Countries/Special Zones in Different Colors

Pros: we are able to fit a trend line in the scatterplot and there's a relationship between suicide rate and total schooling in years.

Cons: Still, there's clearly an outlier in the scatterplot (Luxemburg)

Discussion:

- According to the trend line, in developed countries, suicide rate goes down as total schooling in years goes up.
- Finland has the highest median suicide rate and its median total schooling in years is 3.52.
- Most European countries can fit into the trend line.
- I wonder whether the graph looks very different if I have used the average value instead of the median.

Iteration 3 for Question 2.a:



Chart Type: Scatterplot

Variables encoding:

- Average Suicide Rate (X Axis)
- Average Total Schooling in Years (Y Axis)
- Mark: Selected Developed Countries/Special Zones in Different Colors

Pros: we are able to fit a trend line in the scatterplot and there's a relationship between suicide rate and total schooling in years.

Cons: The average value doesn't visualize potential outliers in the dataset.

Discussion:

- According to the trend line, in developed countries, suicide rate goes down as total schooling in years goes up.
- Finland has the highest average suicide rate and its average total schooling in years is 4.198.
- The UK has the lowest average suicide rate and its average total schooling in years is 5.857.
- Most countries can fit into the trend line. However, I want to know the standard deviation of the dataset so I can determine how well the dataset fits the trend line overall.
- Need to do a similar visualization using median values.

Iteration 4 for Question 2.b:



Chart Type: Scatterplot

Variables encoding:

- Median Suicide Rate (X Axis)
- Median Total Schooling in Years (Y Axis)
- Mark: Selected Developing Countries in Different Colors

Pros: The scatterplot is very clear.

Cons: The trend line generated does not fit all the data really well.

Discussion:

- According to the trend line, in developed countries, suicide rate goes up as total schooling in years goes up.
- Turkey has the lowest median suicide rate and its median total schooling in years is 0.807.
- Russia has the highest median suicide rate and its median total schooling in years is 3.025.
- Most countries cannot fit into the current trend line. We might want to take a look at the standard deviation of this dataset.
- Need to do a similar visualization using average values.

Iteration 5 for Question 2.b:

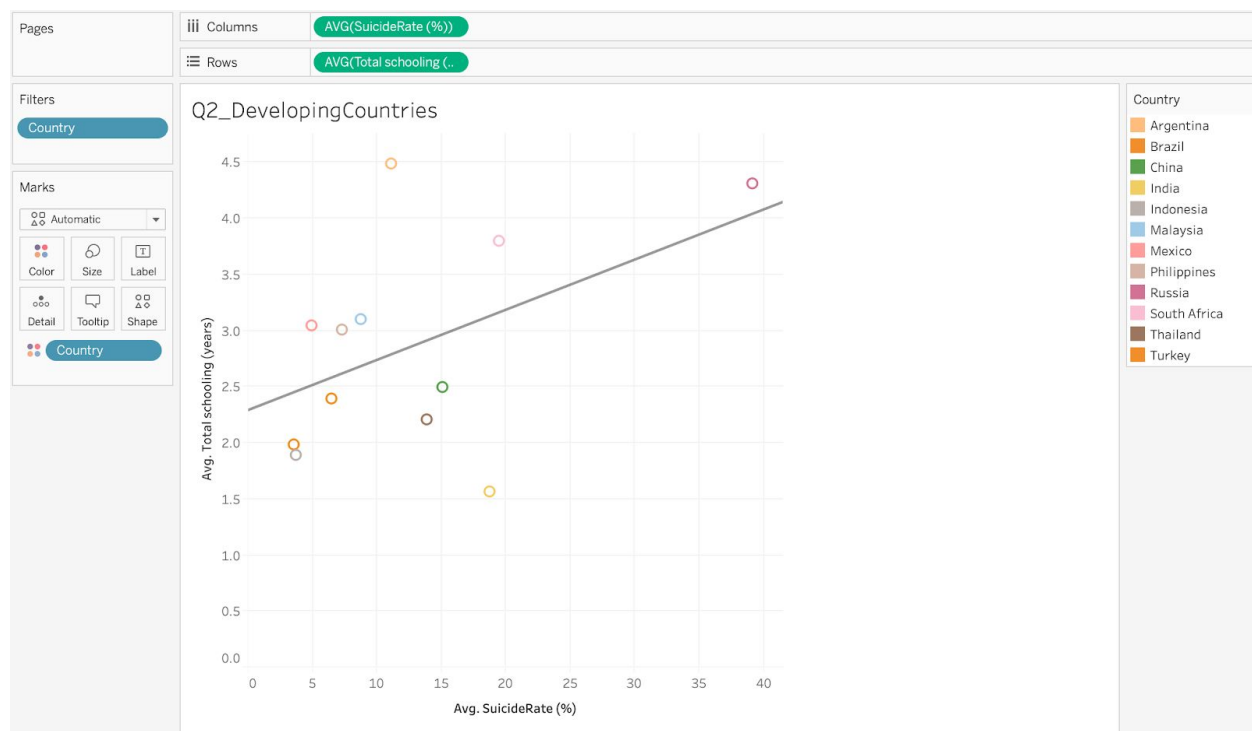


Chart Type: Scatterplot

Variables encoding:

- Average Suicide Rate (X Axis)
- Average Total Schooling in Years (Y Axis)
- Mark: Selected Developing Countries in Different Colors

Pros: The scatterplot is very clear.

Cons: The trend line generated fit the data slightly better than the trend line generated in the previous iteration.

Discussion:

- According to the trend line, in developed countries, suicide rate goes up as total schooling in years goes up.
- Turkey has the lowest average suicide rate and its average total schooling in years is 1.98.
- Russia has the highest median suicide rate and its average total schooling in years is 4.31.
- There are a couple of outliers in the data: Indian and Argentina.

Iteration 6 for Question 2.c:



Chart Type: Scatterplot

Variables encoding:

- Average Total Schooling in Years (X Axis)
- Average Life Expectancy in Years (Y Axis)
- Mark: all countries in different colors

Pros: the scatterplot is very clear. There's a clear relationship between average total schooling and life expectancy.

Cons: N/A

Discussion:

- There is a correlation between average total schooling and average life expectancy. As average total schooling in years goes up, average life expectancy goes up.
- However, this pattern might not remain the same for countries in all regions. Separation of the visualization into regions is needed for the next iteration.

Iteration 7 for Question 2.c:



Chart Type: Scatterplot

Variables encoding:

- Average Total Schooling in Years (X Axis)
- Average Life Expectancy in Years (Y Axis)
- Regions: side-by-side comparison
- Mark: all countries in different colors across regions

Pros: the scatterplot is very clear. There's a clear relationship between average total schooling and life expectancy.

Cons: N/A

Discussion:

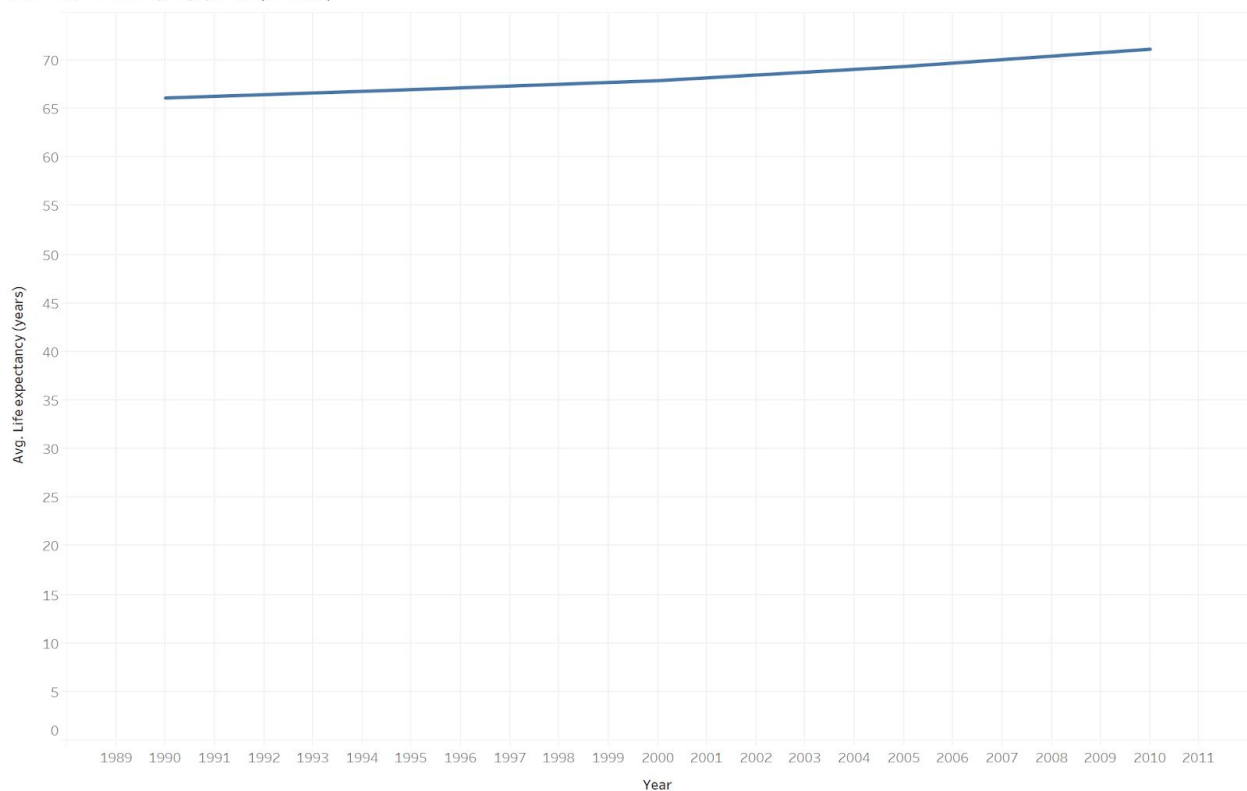
- There is a correlation between average total schooling and average life expectancy.

- The trend line varies based on regions. For Advanced Economies, Asia and the Pacific, Latin America and the Caribbean, Middle East and North Africa, and Sub-Saharan Africa regions, as average total schooling in years goes up, average life expectancy goes up. For Eastern Europe, as average total schooling in years goes up, average life expectancy goes down, which is quite different from our expectation.

Question 3: Is it true that at a global level, the average life expectancy has increased. For countries with higher rates of change in education attainment over a certain time period, the increase in the average life expectancies of these countries is larger during the same time period.

Iteration #1

Life expectancy by year (world)



The trend of average of Life expectancy (years) for Year.

Chart Type: line chart

Variables (encoding):

- Year (X Axis)
- Average Life expectancy (Y Axis)

Pros: Very high level chart to see a birds eye view of the problem at global level. Life expectancy has increased from about 66 years in 1990 to 71 years in 2010.

Cons: Chart is very general and does not tell details

Discussion:

Chart confirms the hypothesis that life expectancy at global level increases with time. However while exploring the data, it appears that the question had two independent parts:

- a) At a global level, the average life expectancy has increased. Is this true at global, regional and country level?
- b) For countries with higher rates of change in education attainment over a certain time period, the increase in the average life expectancies of these countries is larger during the same time period

Updated questions after Iteration #1:

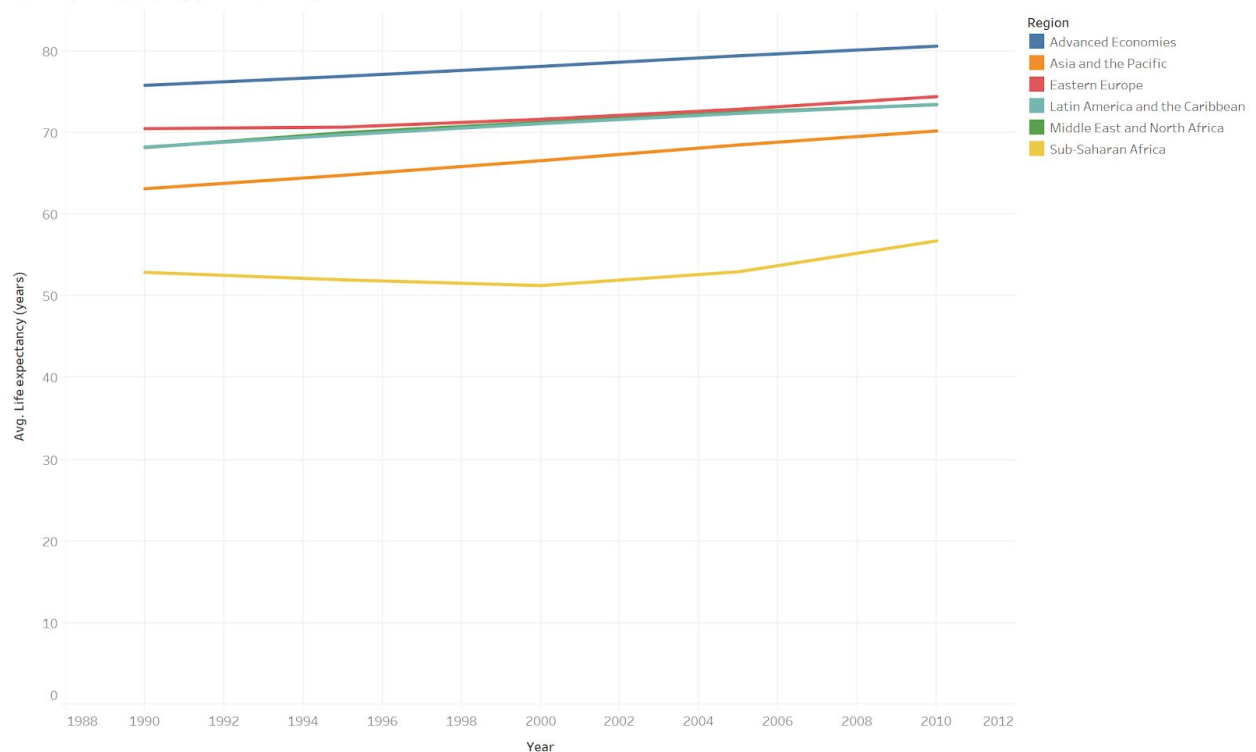
3.a At a global level, the average life expectancy has increased. Is this true at global, regional and country level?

3.b For countries with higher rates of change in education attainment over a certain time period, the increase in the average life expectancies of these countries is larger during the same time period

Question 3.a At a global level, the average life expectancy has increased. Is this true at global, regional and country level?

Iteration #2

Life expectancy by year and region



The trend of average of Life expectancy (years) for Year. Color shows details about Region.

Chart Type: line chart

Variables (encoding):

- Year (X Axis)
- Average Life expectancy (Y Axis)
- Region (color)

Pros: Chart starts to provide more granular data; since there are six regions, color works as a encoding method

Cons: Still high level and there is a need to go down to country level

Discussion: Looking at regional trends in life expectancy, it seems most regions have grown between 1990 and 2010, except Sub-Saharan Africa, which declined between 1990-2000 for and then increased from 2000-2010

Iteration #3

Life expectancy by year and country

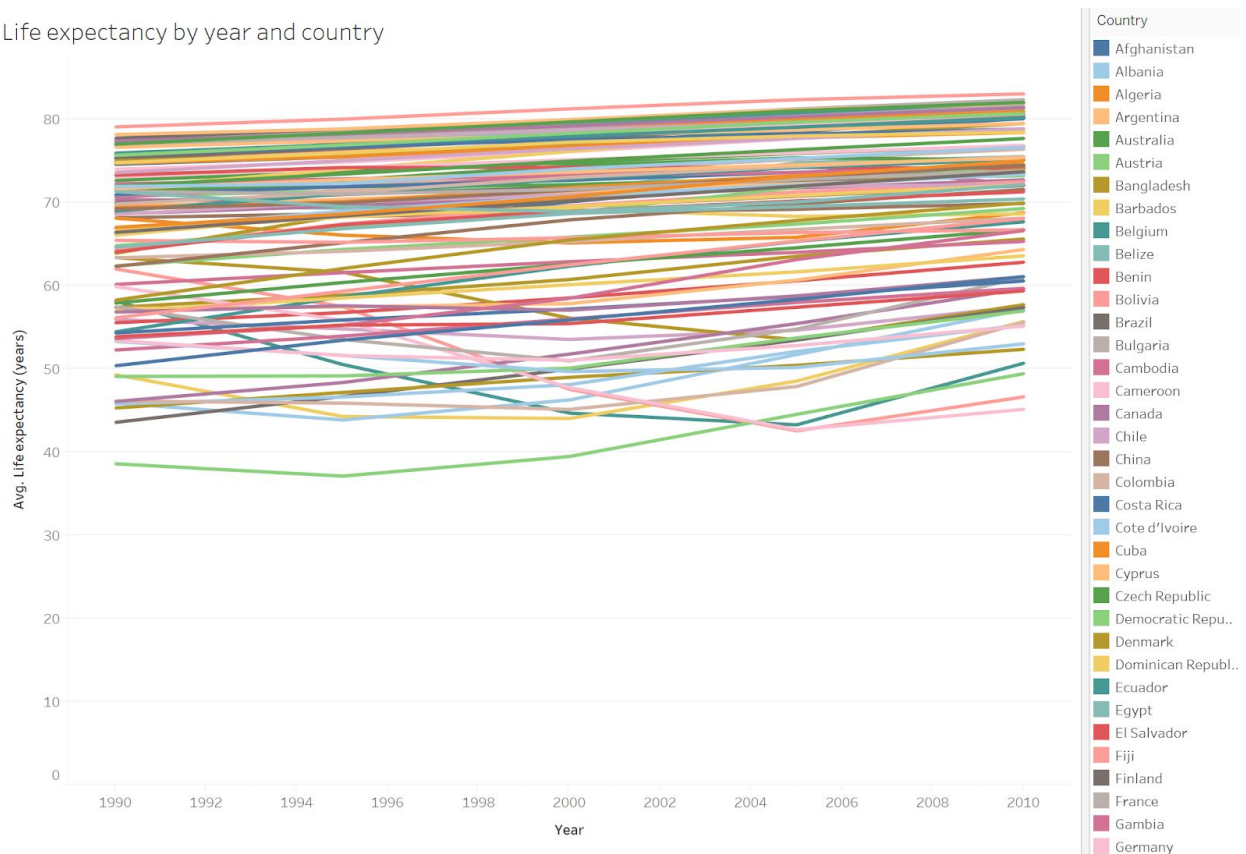


Chart Type: line chart

Variables (encoding):

- Year (X Axis)
- Average Life expectancy (Y Axis)
- Country (color)

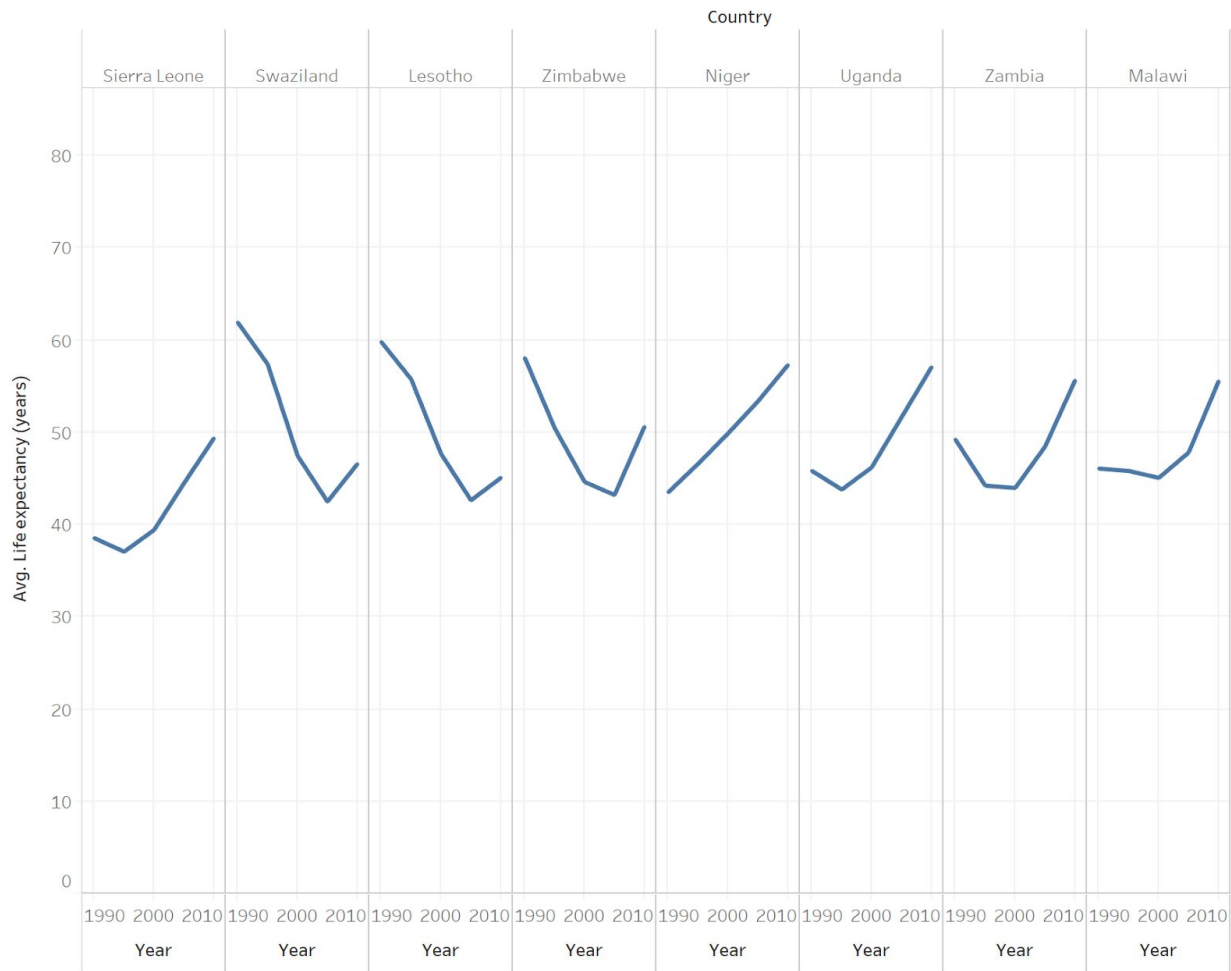
Pros: Chart provides very granular data at country level

Cons: Impossible to get insight as the data is very cluttered, as there are 109 countries

Discussion: Need to find an alternate method of encoding as this method of using color is hindering the analysis

Iteration #4

Life expectancy by year and region



Life expectancy by year and region

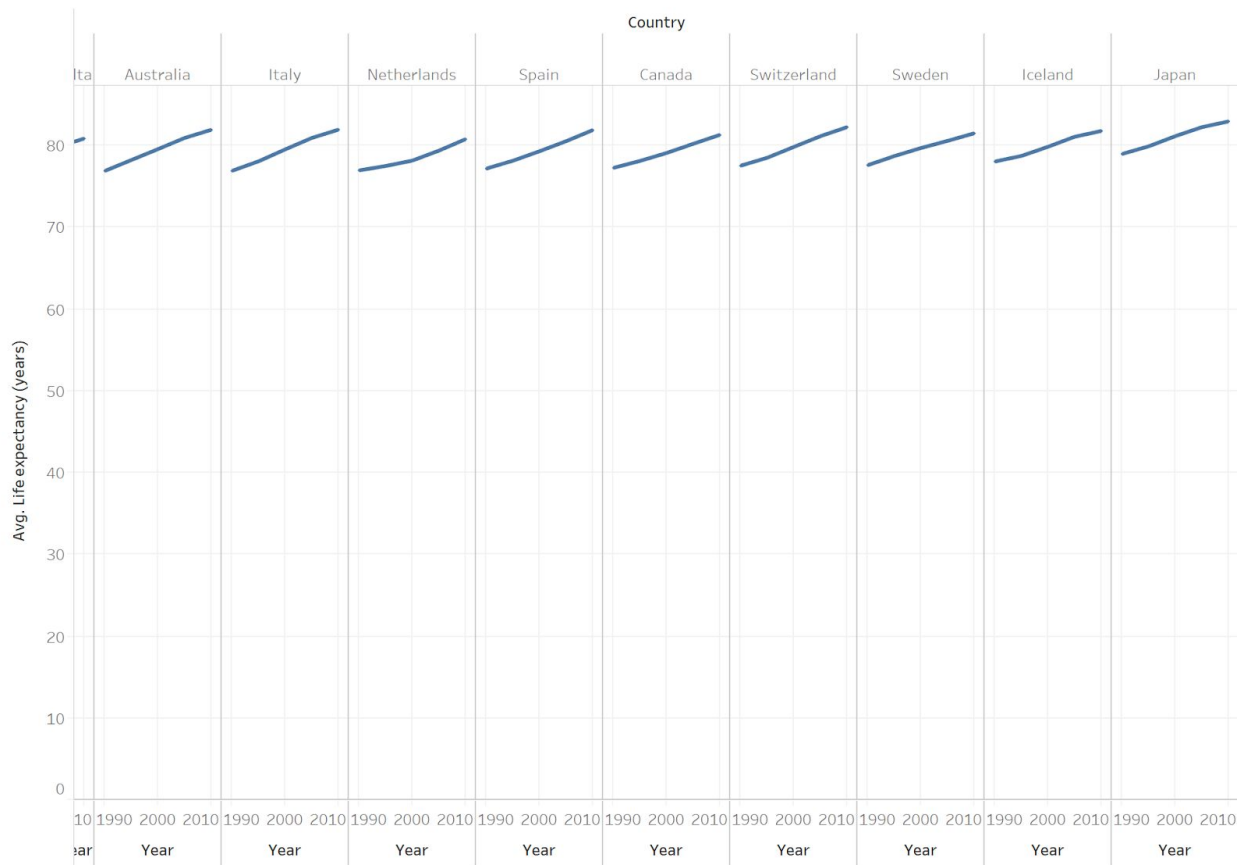


Chart Type: line chart

Variables (encoding):

- Year (X Axis)
- Average Life expectancy (Y Axis)
- Country (space separated)

Pros: Chart provides very granular data at country level

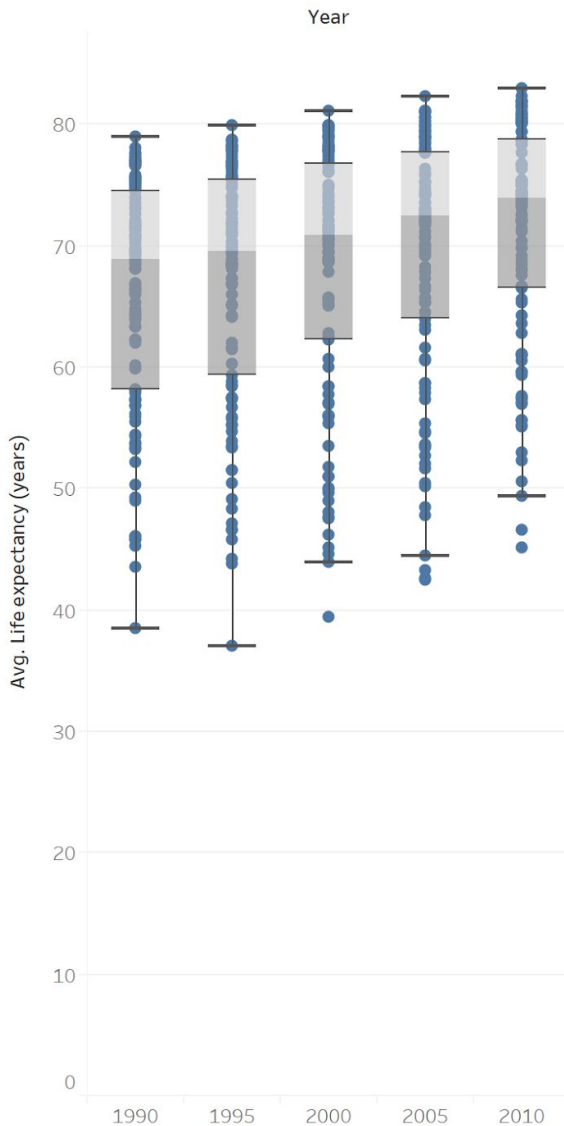
Cons: This chart, albeit long provides a very clear picture for each country and overall patterns

Discussion: This chart shows very clearly how african countries not only have low life expectancy, but it fell from 1990 to 2000. This was unexpected as this differs from the global trends. Countries like Swaziland, Zimbabwe have seen life expectancy to fall till 2005, before rising in 2010. Also as expected Japan has the highest life expectancy. It took a while to find a method in Tableau to be able to chart type that works. One of the challenges in Tableau is that it is very hard to customize visuals. You get what Tableau finds as the best way to visualize the data. I was able to generate the chart after couple of variable changes to rows/columns. Also it was helpful to sort countries by minimum life expectancy.

3.a.1 Additional question posed by this data: What is the distribution of average life expectancy?

Iteration #5

Average Life expectancy by year
(world)



Average of Life expectancy (years) for each Year Year. Details are shown for Country.

Chart Type: box and whisker plot

Variables (encoding):

- Average Life expectancy (Y Axis)

- Year (X Axis)

Pros: Chart enabled comparison of distribution of life expectancy over the years

Cons: Needs understanding of box and whisker plot to gain insight

Discussion: This chart shows how we live in a world where some countries like Sierra Leone have a life expectancy of 49 years and Japan has a life expectancy of 83 (year 2010). This wide gap was unexpected!

Question 3.b: For countries with higher rates of change in education attainment over a certain time period, the increase in the average life expectancies of these countries is larger during the same time period

Iteration #6

Rewording:

Rates of change in education attainment is correlated with the rate of change in life expectancy by countries over a time period

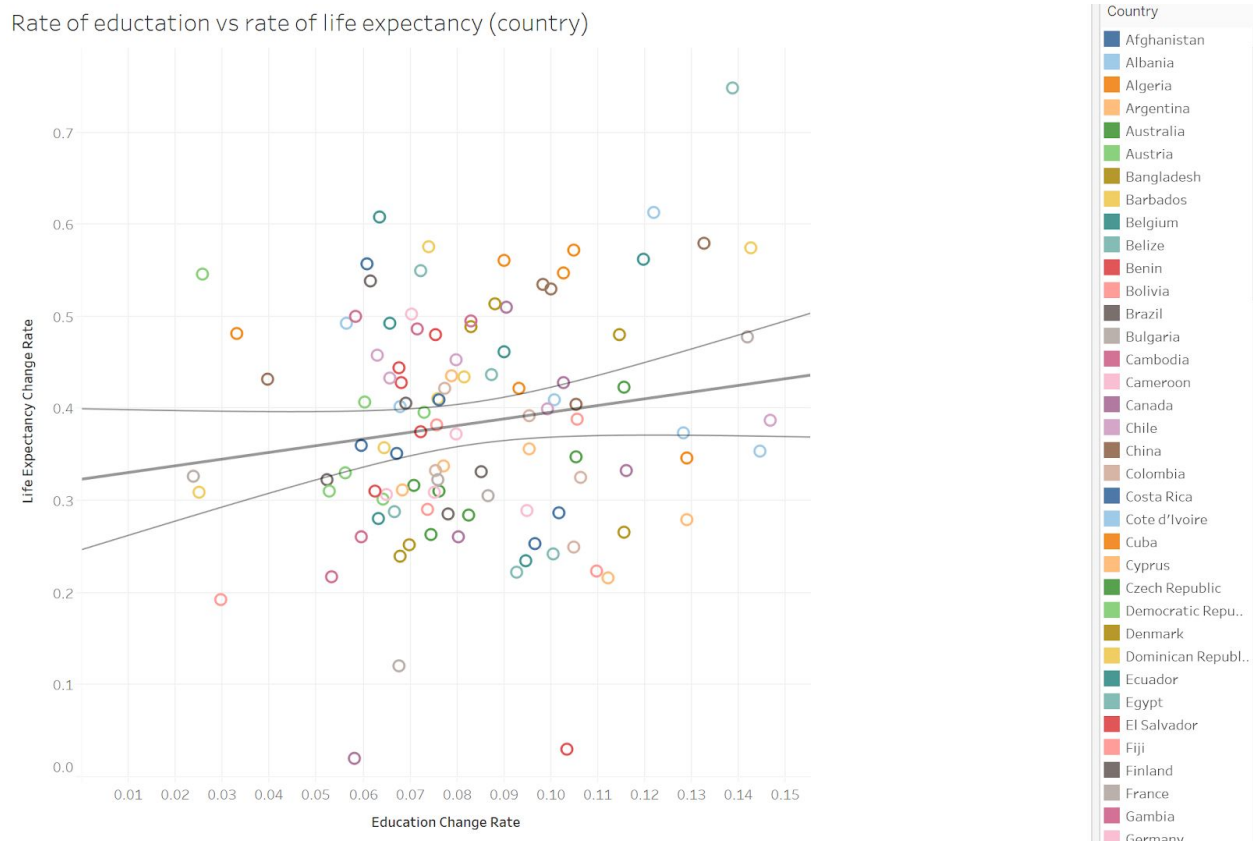


Chart Type: scatter plot

Variables (encoding):

- Life expectancy change rate (Y Axis)
- Education change rate (X Axis)
- Country (color)

Pros: Three variables were encoded well

Cons: since there are 111 countries, harder to see individual countries

Discussion: there is a slight positive correlation between education rate change and life expectancy rate change. There are a lot of outliers.

Question 4: Asia-Pacific (excluding Australia and New Zealand), Africa and South America have higher rates of changes in education attainment compared to North America and Europe

Since the dataset has slightly different names for regions, than the question, reworded the question to:

Question 4 (updated): Asia-Pacific, Africa and Latin America have higher rates of changes in education attainment compared to Advanced economies (North America, West europe etc)

Rate of education vs rate of life expectancy (region)

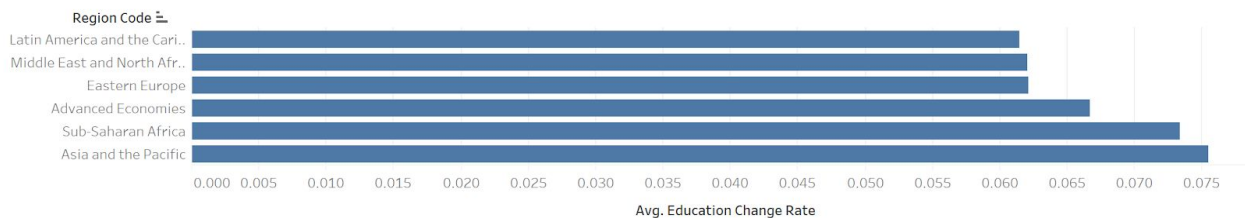


Chart Type: scatter plot

Variables (encoding):

- Education change rate (X Axis)
- Region (Y axis)

Pros: Two variables were encoded well

Cons: none

Discussion: Hypothesis is partly correct as Asia-Pacific and Africa does have a higher rate of education attainment compared to Advanced economies. However, Latin America has lower, in-fact least change in education attainment compared to all regions.

Final Visualization and Question



We uncovered a lot of interesting information during our visual exploration. Among all the visualizations we created, we picked the visualization of the relationship between average total schooling in years(education attainment) and average life expectancy. The visualization has a side-by-side comparison among different regions in the world. From this visualization, we want to ask **a finalized question**:

What are the factors that might affect the correlation between average life expectancy and education attainment among regions? – given that the direction of the trend lines varies based on regions.

Visualization tool and Data cleanup method

We used following tools for data

- Data cleanup tool:
 - Python - We used python to do a merge of the education, life expectancy and suicide rate datasets.
 - Tableau prep - We created a flow in Tableau prep to clean the data
- Visualization tool: Tableau
- Ad-hoc data exploration: Excel

Python: We used Anaconda distribution of Python and used Jupyter notebook for development. Education dataset was in stata format and the pandas IO library allowed conversion to csv. Education data was merged with life expectancy and suicide rate on country and year. Education data is in 5 year intervals and since we did a left outer join, our final dataset has data in 5 year intervals.

Tableau Prep: Tableau prep is a great tool for data cleanup and we were able to develop a flow to clean the data. We generated these outputs:

- a) Education + life expectancy + suicide rate dataset (with nulls)
- b) Education + life expectancy + suicide rate dataset (with no nulls)
- c) Education + life expectancy (no nulls)

However there are some shortcomings of Tableau Prep:

- a) If the app crashes, it does not recover your unsaved flow. So one needs to keep saving the flow while making progress
- b) It is not obvious that an output step is required. We had to research online on finding ways for the prep to save output as a csv
- c) Tableau's support for datetime can be better. Our dataset contains year as a column. When we converted it to date, it gave a weird date instead of 1/1/<year>

Tableau desktop: We used Tableau, and while the tool is powerful, it has these shortcomings:

- a) It is hard to control how you want your visualization to render. Tableau makes a lot of rendering decisions automatically. So it needs a lot of trial and error to get to a right visual
- b) Built-in default trend lines are susceptible to outliers

Conclusion:

Overall we discovered that the data partly agreed with our hypothesis and partly not.

- Education attainment (total schooling in years) does have an upward curve globally, regionally, and nationally overtime.
- Regions and countries matter when it comes to the correlation between education attainment and life expectancy.
- For iteration 4 for Question 2.b, the trend line generated by Tableau does not fit the dataset well. We should find a better way to generate a trend line that can fit the dataset for our future work.
- Worldwide, life expectancy has increased over time. However, looking at overall distribution of life expectancy by year, it is surprising to see a wide gap. In 2010, Sierra Leone had a life expectancy of 49 years and Japan has a life expectancy of 83.
- Looking at time series data for life expectancy by country, shows an interesting trend. Some countries such as have a “hockey-stick” pattern where the life expectancy decreased between 1990-2000 and then increased till 2010. This is most evident in African countries such as Sierra Leone, Swaziland, Uganda. Sierra Leone had a civil war in the 1990s and that can explain some of the decrease in life expectancy.