Education Data Visualization Exploratory Analysis

Anqi Cao, Shiyao Chen

Table of Contents

Introduction

Education, as a fundamental resource both for individual and social development, has been a field studied and researched by innumerable scholars and professionals. With the rich data available on this topic, we are curious about how different dimensions of education attainment have changed in recent history at a global level and their correlational impacts on other facets of development. Considering that both life expectancy and education attainment constitute the two important aspects of the Human Development Index published regularly by the UN's Development Programme, we believe there should be a correlation between the two worth exploring. Furthermore, given that education has been one of the social determinants in mental health according to existing studies of regional data, we'd love to dig into how mental health conditions have fluctuated longitudinally with the changes in education attainment for different countries. To this end, for the purpose of simplifying the process of data collection and aggregation, we have settled on suicide rate, which is a major indicator of a population's mental health.

We hope that by visualizing datasets pertaining to education attainment, life expectancy and suicide rate, we can provide researchers, students and educational professionals with an overview of the development of education by region and time period, and what role education has played in shaping some of the major facets of human development such as physical and mental health.

Datasets

Dataset #1: Educational Attainment (Primary dataset)

Name: Barro-Lee Education Attainment Data

Description: The Barro-Lee Data set provides education attainment data for 146 countries from 1820 to 2010 disaggregated by sex and 5-year intervals. It also provides information about the distribution of educational attainment of the adult population at 7 levels of schooling— no formal education, incomplete primary, complete primary, lower secondary, upper secondary, incomplete tertiary, and complete tertiary. Average years of schooling at all levels—primary, secondary, and tertiary—are also measured for each country and for regions in the world.

Size: 12987 rows x 23 columns

Dimensions and data types:

- Region code (nominal)
- Country name (nominal)
- Year (interval)
- Age (ratio)
- Sex (nominal)
- Primary adjusted enrollment ratio (%) (ratio)
- Secondary adjusted enrollment ratio (%) (ratio)
- Tertiary adjusted enrollment ratio (%) (ratio)
- Percentage of no schooling (ratio)
- Percentage of primary (ratio)
- Percentage of primary complete (ratio)
- Percentage of secondary (ratio)
- Percentage of secondary complete (ratio)
- Percentage of tertiary (ratio)
- Percentage of tertiary complete (ratio)
- Years of schooling (ratio)
- Years of primary schooling (ratio)
- Years of secondary schooling (ratio)
- Years of Tertiary Schooling; (ratio)
- Human capital, population aged 15-64 years (ratio)
- Alternative human capital, population aged 15-64 years (ratio)
- Population (thousands) (ratio)

Source: http://www.barrolee.com/v

How we intend to use it: Play around with the dataset to generate a geospatial graph that allows the audience to have an overview of the rate of education attainment growth globally, filter by time period and type of education attainment and zoom in on the details associated with each geographical region.

Dataset #2: Life Expectancy (Secondary dataset)

Name: Time series data of differences in life expectancy across the world

Description: The dataset aggregate the life expectancies of countries from 1543 to 2019 with gaps for certain countries and certain time periods

Size: 19028 rows x 4 columns

Dimensions and data types:

- Country (nominal)
- country code (nominal)
- Year (interval)
- life expectancy (ratio)

Source: https://ourworldindata.org/life-expectancy

How we intend to use it: Correlate the time series data of life expectancy by country with the corresponding country's data of education attainment

Dataset #3: Suicide Rate (Secondary dataset)

Description: The dataset provides share of suicide deaths per 100,000 individuals from 1990 to 2017 by country

Size: 6469 rows x 4 columns

Dimensions and data types:

- Country (nominal)
- country code (nominal)
- Year (interval)
- % of deaths from suicide (ratio)

Source: http://ghdx.healthdata.org/gbd-results-tool

How we intend to use it: Correlate the time series data of suicide % by country with the corresponding country's data of education attainment

Note: All metrics are age-standardized to allow comparisons between countreis and over time.

We have merged all three datasets by joining them on country and year as keys. This process required country name cleanup and was done in python. Merged data is available here:
https://github.com/edu-infoviz/edu/blob/master/data/edu_merged_data.csv

Cleanup and merge script is available here:
https://github.com/edu-infoviz/edu/blob/master/edu/EduProject.ipynb

Initial Research Questions

Initial Question 1:

How do countries compare in terms of changes in “years of no schooling” and “percentages of tertiary complete” from 1950 to 2010?

- **Rationale:** It is generally assumed that literacy rate and the size of the educated population have been rising over the past decades globally, but it is worth investigating how other dimensions of educational attainment have changed for different regions and see how changes in these dimensions may or may not follow the same trend. In this case, we have selected change in “years of no schooling” as a measure that implies change in literacy rate, and show how higher education attainment as represented by the percentage of completed tertiary degrees has varied across regions and changed over time.
- **Implications for users:** Researchers and students interested in digging into the nuances of educational attainment across time and space will benefit from the visualization.
- **Refinements:** In order to have a leaner view, we’ve decided to use region codes instead of countries as nominal variables on the x axis, specifically the Advanced Economies, Asia and the Pacific, Latin America and the Carribean, Middle East and North Africa, and Sub-Saharan Africa.
-

Initial Question 2:

Do different dimensions of education attainment correlate with life expectancy, and if so, in what regions?

- **Rationale:** It is reasonable to synthesize that education attainment positively correlates to life expectancy. People that get educated have more knowledge and care about how to live a healthy life. Also healthcare develops better, with more education efforts and talents injected into this industry. This positive relationship might not exist in all countries, which is worth exploring.
- **Implications for users:** Researchers, policy makers and scholars interested in the implications of educational attainment in the field of population health can explore this visualization

- **Refinements:** To operationalize education attainment, we decided to use “average years of schooling” as the dimension.

Initial Question 3:

Do different dimensions of education attainment correlate with suicide rate, and if so, in what regions?

- **Rationale:** Researches have shown the negative correlation between education attainment and suicide rate for certain regions during specific time periods. It will be interesting to compare data in the same time period across regions, and across time periods in the same region, to see whether the negative correlation still holds.
- **Implications for users:** Researchers, policy makers and scholars interested in the implications of educational attainment in the field of mental health can explore this visualization
- **Refinements:** To operationalize education attainment, we decided to use “average years of schooling” as the dimension.

Initial Question 4:

Do different regions have different growth rates in educational attainment and life expectancy?

- **Rationale:** It’s no doubt that the global educational attainment and life expectancy are both growing. But the growth rates might be different. Some developing countries like China and India might experience higher growth rates compared to developed ones, although the overall levels are still below them.
- **Implications for users:**Researchers, policy makers and scholars interested in the implications of educational attainment in the field of population health can explore this visualization
- **Refinements:** To operationalize education attainment, we decided to use “average years of schooling” as the dimension.

Initial Question 5:

Do other attributes of countries have relationships with education level and life expectancy? Any interesting patterns?

- **Rationale:** There are many factors mentioned frequently when comparing different countries, like population. Research has shown many small countries rank in top places in terms of livability. Education level and life expectancy can be a measure or a evidence of it.
- **Implications for users:** Any researchers or public interested in public policy,
- **Refinements:** To operationalize education attainment, we decided to use “average years of schooling” as the dimension.

Visual Exploration Process

Q #1: How do different regions (the Advanced Economies, Asia and the Pacific, Latin America and the Carribean, Middle East and North Africa, and Sub-Saharan Africa) compare in terms of changes in “percentage of no schooling” and “percentage of tertiary complete” from 1950 to 2010?

Question 1: Iteration 1

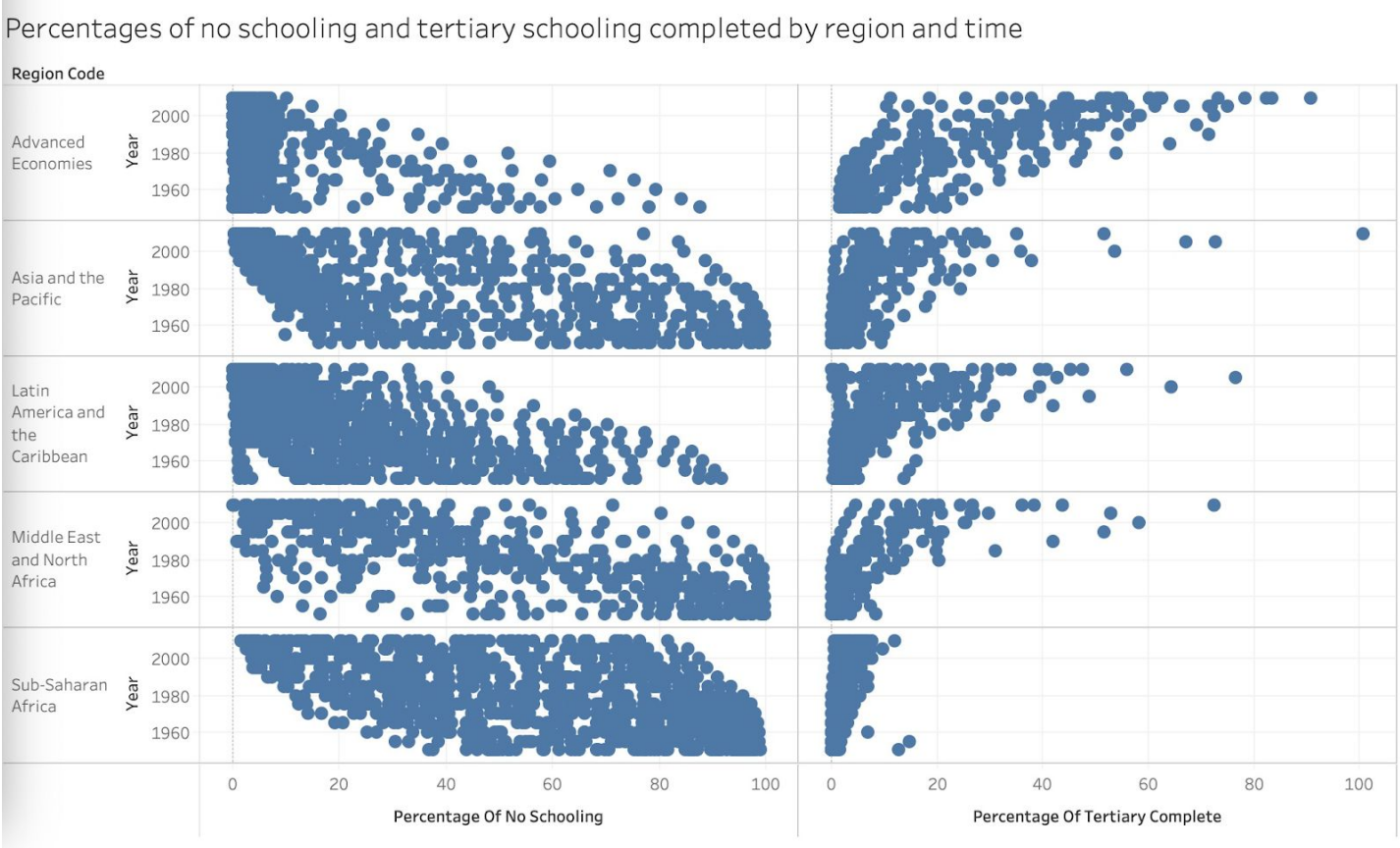


Chart type: Scatter Plot
Variables (encoding):
Region code (Y axis)
Year (Y axis)
Percentage of No Schooling (X axis)
Percentage of Tertiary Complete (X axis)
Country name (mark)

PROS

- Has all the variables displayed and a clear trend can be spotted for each region over time based on

the density of the dot cluster

CONS

- Monochromatic
- Percentage of no schooling and percentage of tertiary complete are divided into two sections along the x axis
- The way the time is sequenced is not intuitive

DISCUSSION

- It'll be better to use two dimensions of one visual encoding mark to represent two variables: percentages of no schooling and tertiary complete, thus combining the two sections into one
- Need to display time in a more intuitive way, such as in an ascending order to offer a better visual experience

Question 1: Iteration 2

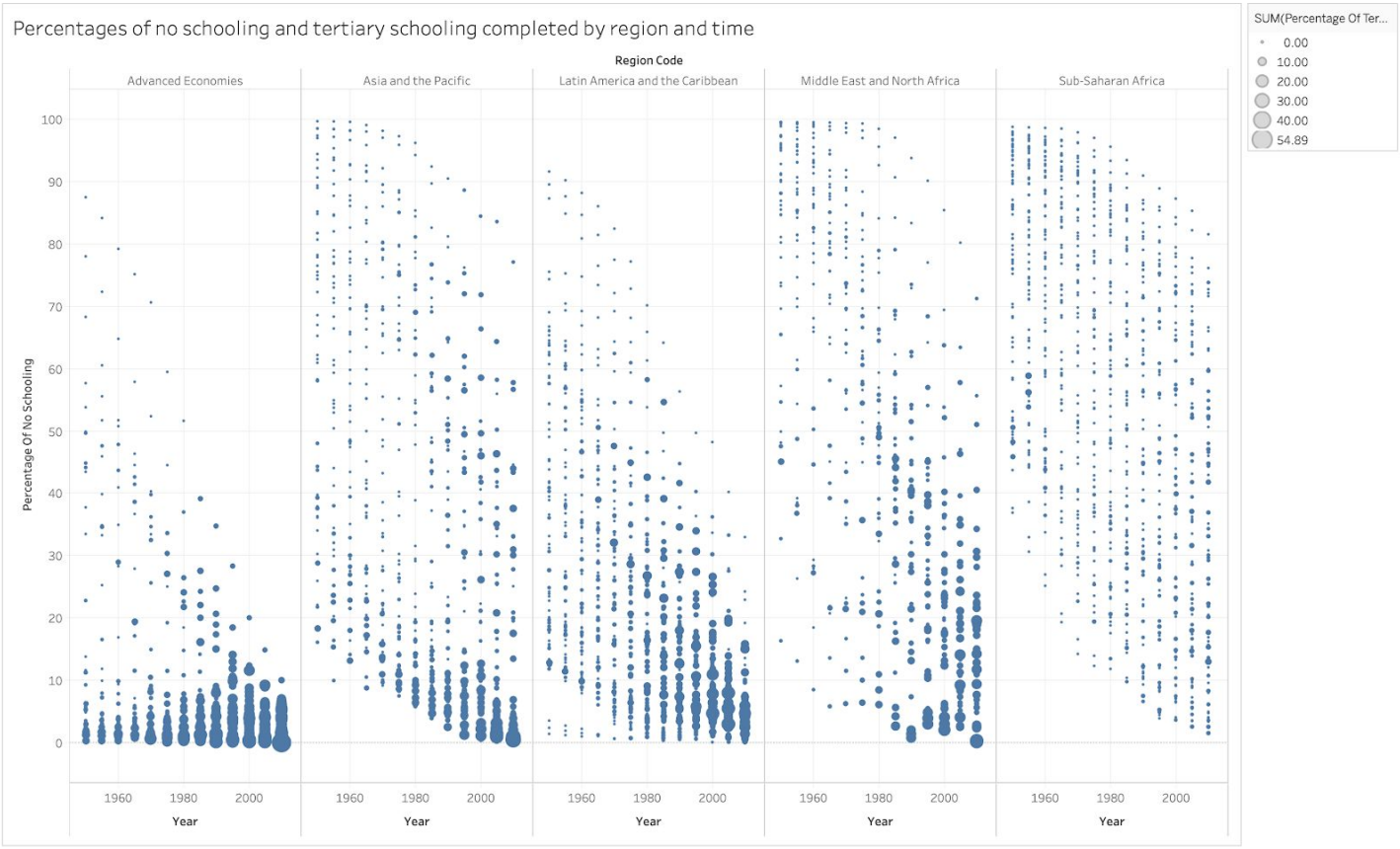


Chart type: Scatter Plot
Variables (encoding):
Region code (Y axis)
Year (Y axis)
Percentage of No Schooling (X axis)
Percentage of Tertiary Complete (mark - size)
Country name (mark - detail)

PROS

- Trends in percentages of no schooling and tertiary complete can be clearly identified in an ascending chronological order by region
- The variation in the size of the dot reduces information cluster and makes the whole graph look cleaner

CONS

- Monochromatic, not visually engaging

DISCUSSION

- It will be interesting to explore other possible views of data in addition to scatter plot, such as a geospatial representation. Currently the data is disaggregated by region, which has no assigned geographical role, but how will the map look like if we map the data geospatially by each country?

Question 1: Iteration 3

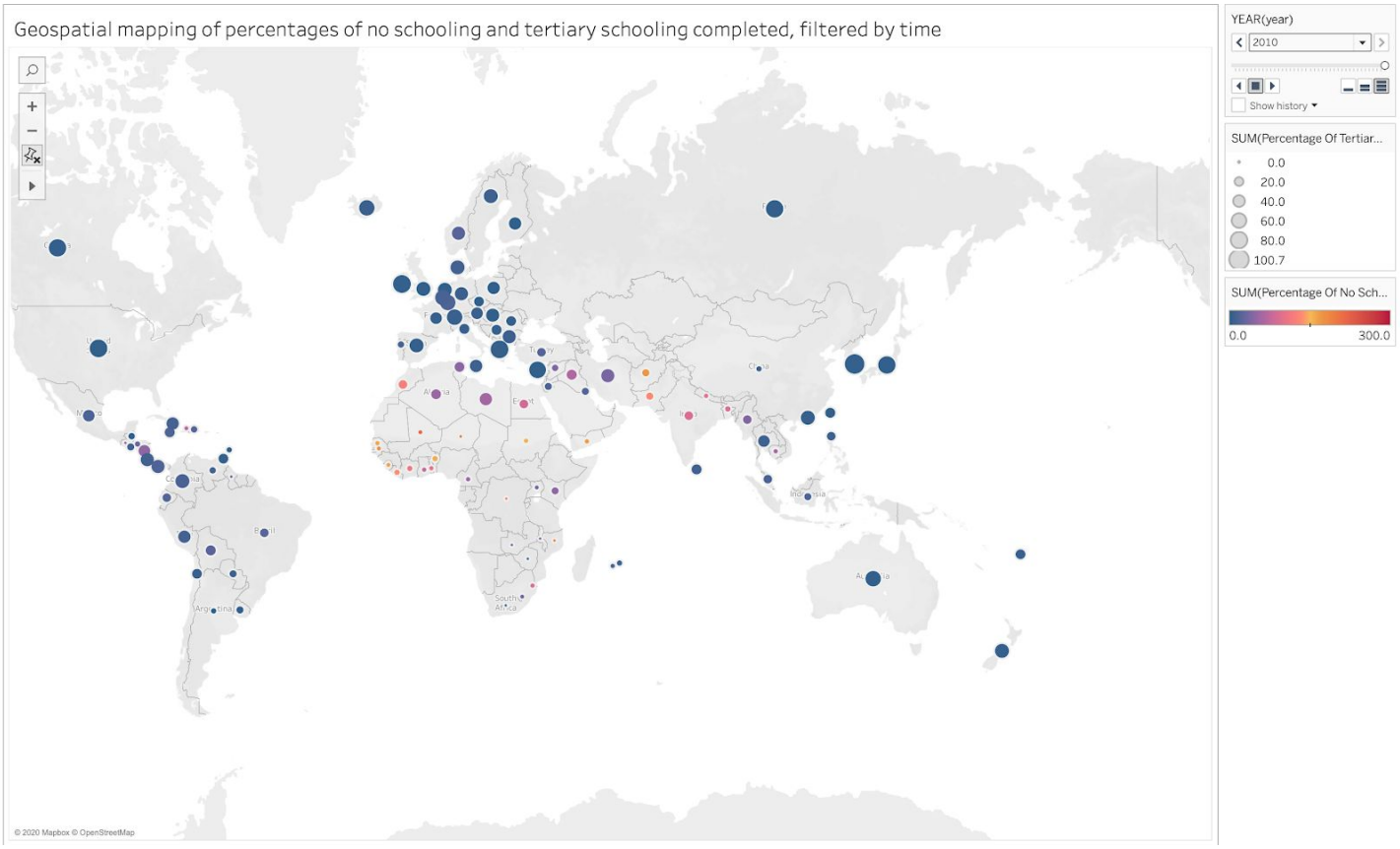


Chart type: Map
Variables (encoding):
Country (longitude and latitude)
Year (pages)
Percentage of No Schooling (color)
Percentage of Tertiary Complete (size)
Country name (mark - detail)

PROS

- The data map displays the geospatial relationships between countries in a visually forward way, eliminating the need to use region code
- The sharp contrast in dot size makes it clear how the disparity in tertiary education attainment plays out in different geographical regions
- Time as “pages” allows users to conveniently filter by the year they are interested in, and offers the option for them to view the longitudinal changes in a continuous, animated way

CONS

- The use of blue - red diverging gradience to encode a numerical variable (percentage of no schooling) is not as intuitive as position along the y axis

- Cannot display time-series data

DISCUSSION

- It does a good job mapping global education information at a particular time in history, but it requires a learning process where users familiarize themselves with the visual encodings of data dimensions to have a better grasp of what the changes in size and color imply

Q #2: Do different dimensions of education attainment correlate with life expectancy, and if so, in what regions?

Question 2: Iteration 1

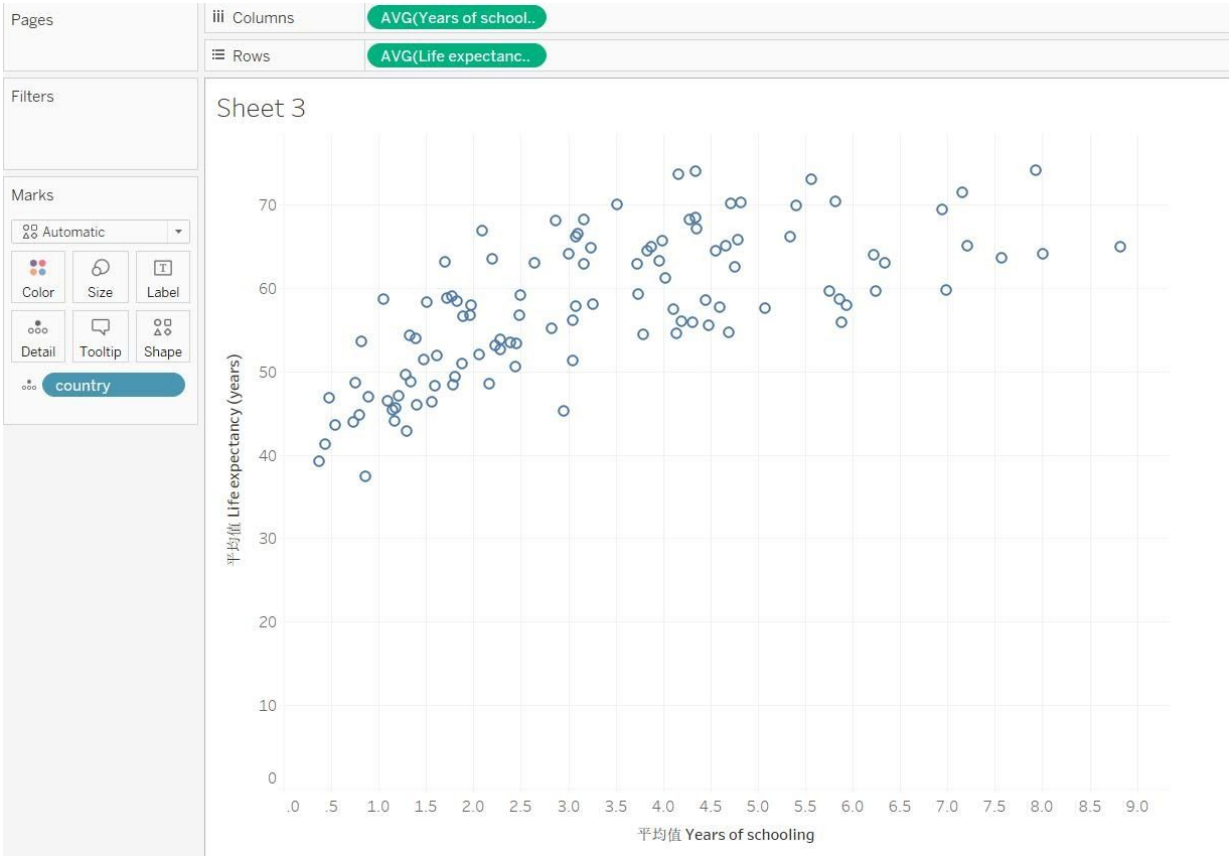


Chart type: Scatter Plot
Variables (encoding):
Avg Life expectancy (Y axis)
Avg Years of schooling (X axis)
Country (Position)

PROS

- Has all the variables displayed and a clear trend can be spotted

CONS

- Cannot compare between regions
- Cannot compare the change with time

DISCUSSION

- Maybe use color to show all countries in several regions

Alternatives of iteration 1

The previous x axis, Avg years of schooling, was replaced by avg percentage of primary, avg percentage of secondary, avg percentage of tertiary. These three scatter plots also show a similar trend with iteration 1: population's life expectancy has a positive correlation with education level to some extent. The countries with the lowest life expectancy have the most underdeveloped education. The countries with the highest life expectancy are not those with the highest education level.

Question 2: Iteration 2

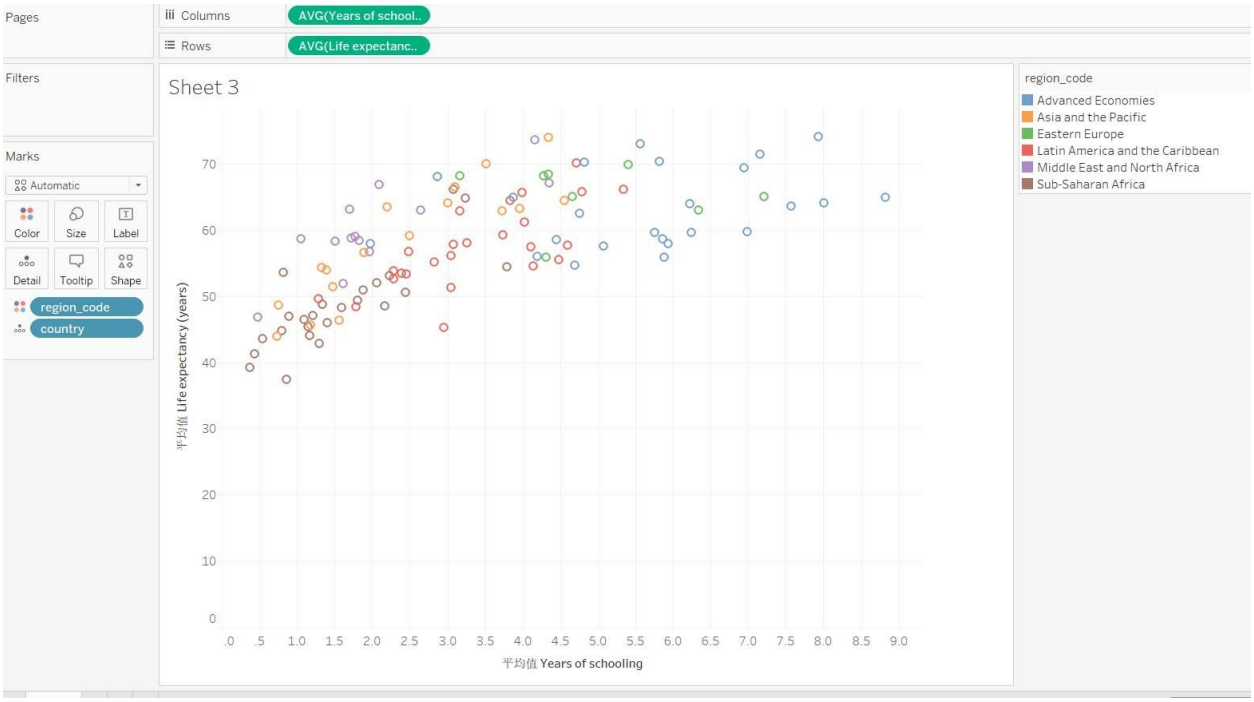


Chart type: Scatter Plot
Variables (encoding):
Avg Life expectancy (Y axis)
Avg Years of schooling (X axis)
Country (Position)
Region (color)

PROS

- Has all the variables displayed and a clear trend can be spotted
- It allows users to prepare different regions

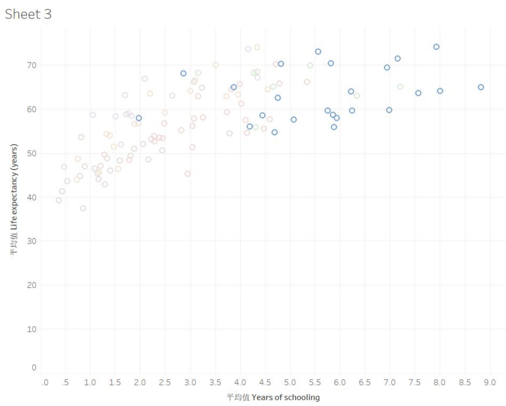
CONS

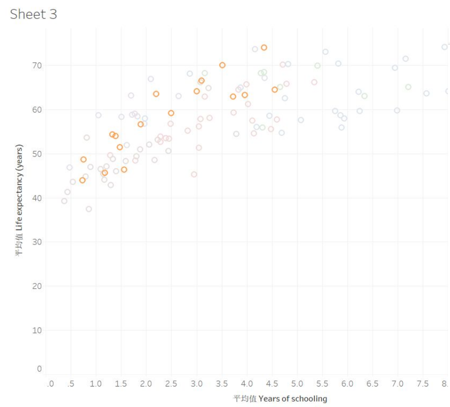
- It's hard to see the trend of each region as the colors are kind of hard to tell when placed together.
- Cannot compare the change with time

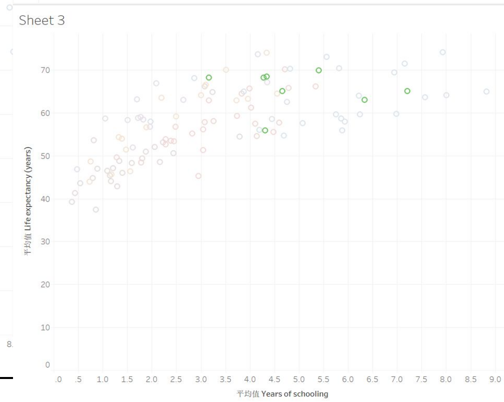
DISCUSSION

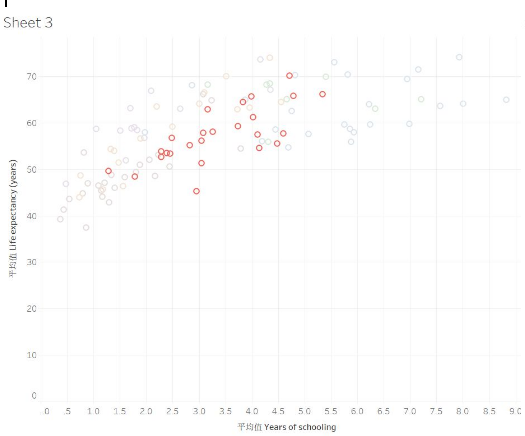
- Maybe use colors with more differences to better compare

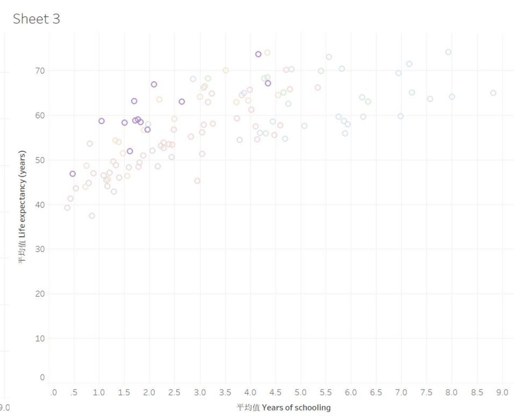
Highlight regions in Iteration 2:

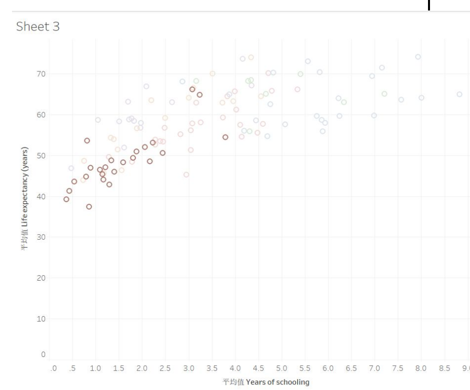












Question 2: Iteration 3

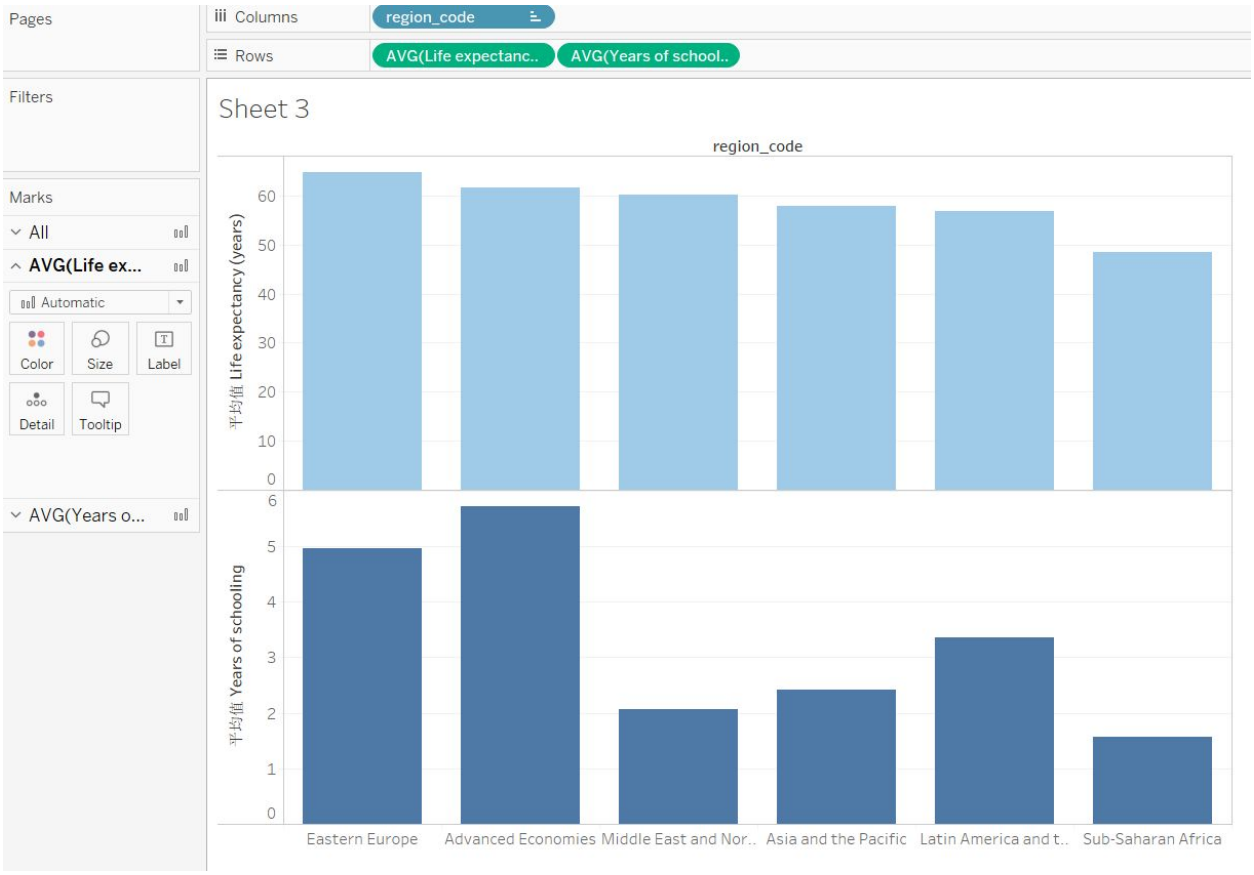


Chart type: Bar Chart
Variables (encoding):
Avg Life expectancy (Y axis)
Avg Years of schooling (Y axis)
Regions (X axis)
Different y axis variables (color)

PROS

- Ranking regions in terms of the two areas clearly
- It allows users to prepare different regions

CONS

- Cannot see what happened in detail in each region
- Cannot compare the change with time
- As the two types of bars are not placed on the same axis, it's kind of hard to compare these two variables together.

DISCUSSION

- Put them on the same x axis.

Question 2: Iteration 4

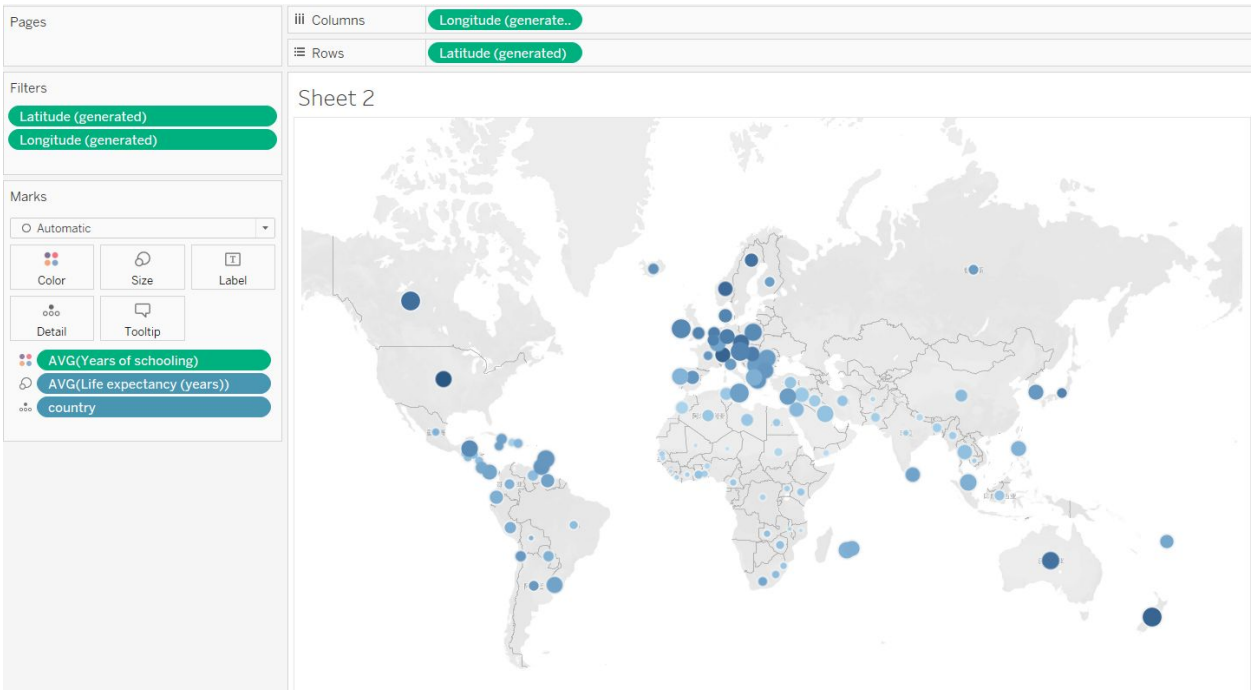


Chart type: Map
Variables (encoding):
Latitude (Y axis)
Longitude (X axis)
Avg Life expectancy (Color)
Avg Years of schooling (Size)
Country (Position)

PROS

- Show several quantitative measurables clearly
- Show geographic positions so users can compare the education and life expectancy between different states

CONS

- First impression is like a density map, which may misleads users that the denser places have higher education levels or life expectancy.
- Due to lack of data of some places, it may misleads users that places without dots are the places without education.
- Cannot see the real numbers directly

DISCUSSION

- Use the shapes of countries themselves as marks, rather than dots.

Q #3: Do different dimensions of education attainment correlate with suicide rate, and if so, in what regions?

Question 3: Iteration 1

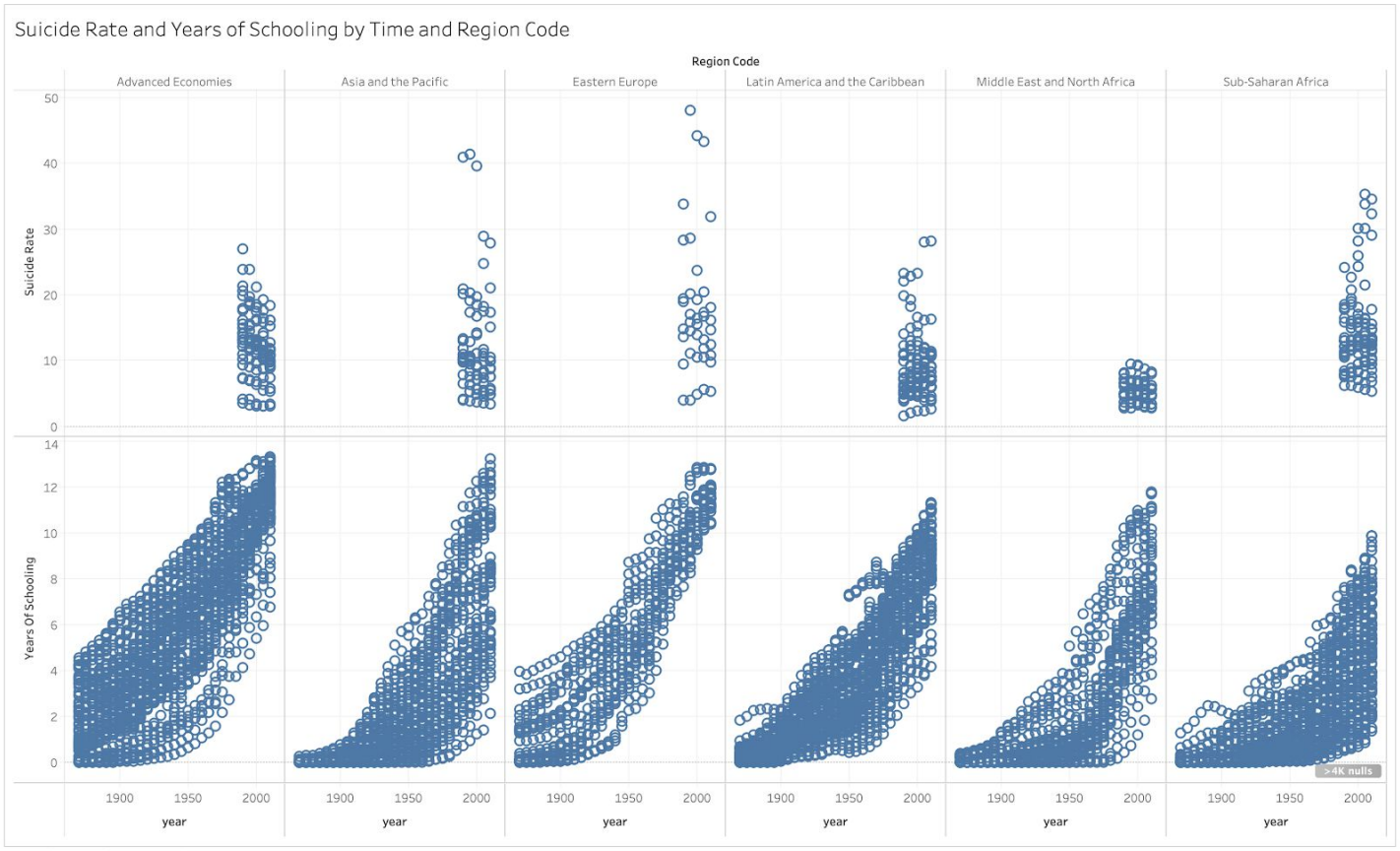


Chart type: Scatter plot
Variables (encoding):
Region code (x axis)
Year (x axis)
Years of schooling (y axis)
Suicide rate (per 100,000 individuals) (y axis)
Country name (mark)

PROS

- A good start for further data scrubbing
- Trends can be identified by looking at the shape of the distribution

CONS

- The time span associated with years of schooling is much wider than that of suicide rate
- Aesthetically unpleasing

DISCUSSION

- To start, we picked “average years of schooling” as the education attainment dimension and see if there’s any correlation between it and suicide rate
- It’s hard to identify any prominent trend of suicide rate for all of the regions, and the way suicide rate and years of schooling are segregated makes it hard to correlate the two. It’ll be better to filter out the time period during which the data on suicide rate is not available.

Question 3: Iteration 2

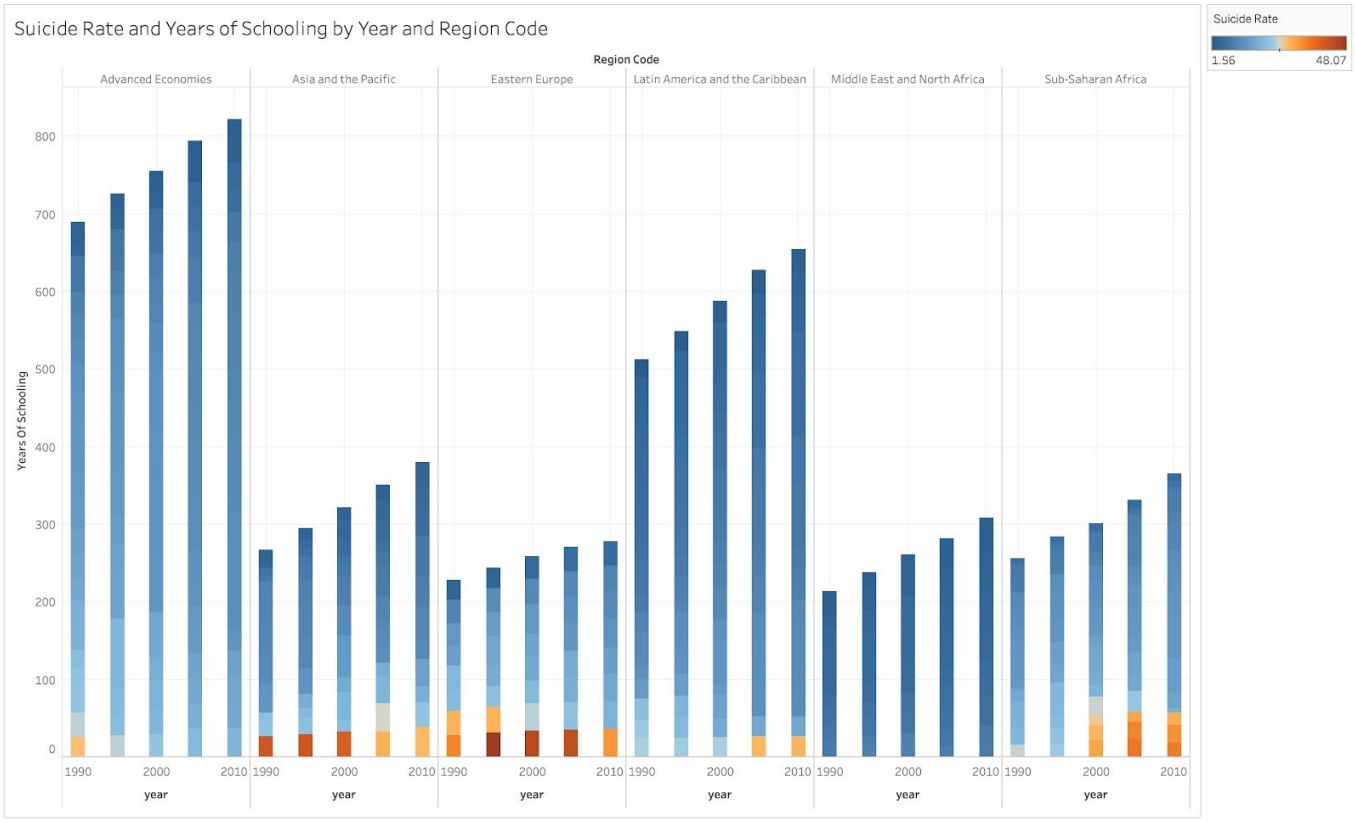


Chart type: Bar Chart
Variables (encoding):
Region code (x axis)
Year (x axis)
Years of schooling (y axis)
Suicide rate (per 100,000 individuals) (color)
Country name (mark)

PROS

- A straightforward representation of the relative avg numbers of years of schooling across time and regions clarifies trends much better
- The bars for average years of schooling are marked by color to show the suicide rate as well, thereby encoding the two variables on the same shape for a convenient study of the correlation between the two.
- The orange-blue diverging gradience allows for a clear observation of the differences in suicide rates

CONS

- It's challenging, if not entirely impossible, to quantify color codes and compare between the color values and hues of different bars, especially when the color values and hues are quite similar to each other. Therefore, it can't tell us how suicide rates have changed over time in a straightforward way.

Only the change in suicide rate for the country with the highest suicide rate (coded by the deepest orange hue) can be observed in a visually forward way.

- Not accessible to people with visual impairment or difficulty in distinguishing between colors

DISCUSSION

- To iterate on the old chart, we performed the following steps:
    - Used bars instead of dots to represent excluded 30 null data in the Sub-Saharan Africa region; filtered out the years for which the suicide rate data isn't available

Question 3: Iteration 3

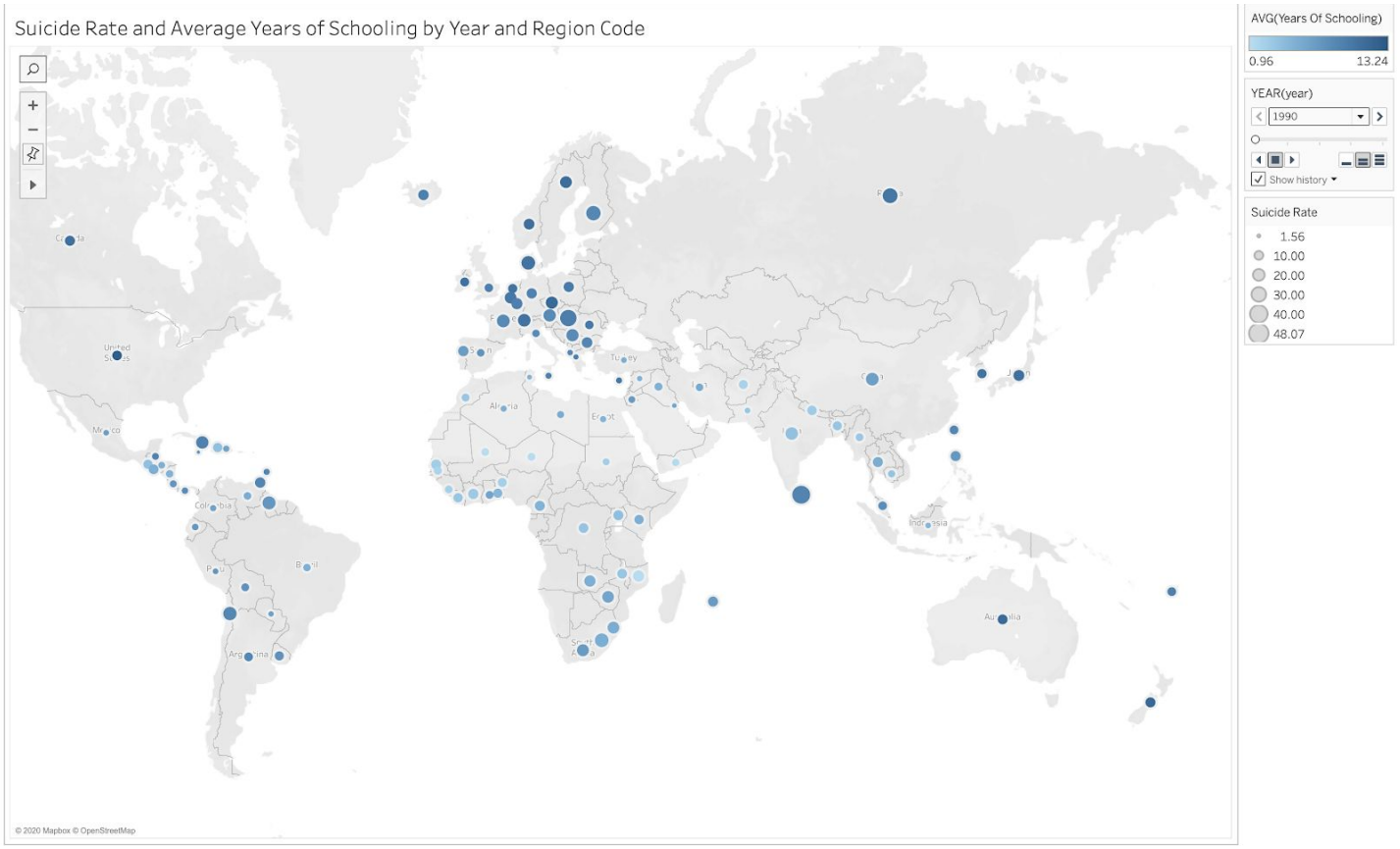


Chart type: Map
Variables (encoding):
Country (longitude and latitude)
Year (pages)
Years of Schooling (continuous color)
Suicide rate (per 100,000) (size)
Country name (mark - detail)

PROS

- Same as the data map under question #1, the data map displays the geospatial relationships between countries in a visually forward way, eliminating the need to use region code
- Since the country data are not aggregated into the same bar for each year, but rather displayed separately based on their geographical locations and viewed by year, it's easier to compare within and between countries their education attainment and suicide rate.

CONS

- Longitudinal trend is hard to be identified, both because of the map's inability to display time series data the same page and the lack of data on suicide rate across a sufficiently large time span.

DISCUSSION

- It seems that there's no clear, consistent direction in terms of how the intensity of hue changes with the size of the circle, making it hard to reach a conclusion of the correlation between suicide rate and education attainment as in years of schooling in this case
- However, this observation doesn't negate the possibility of the correlation between other education variables and suicide rate

Initial Question 4: Do different regions have different growth rates in educational attainment and life expectancy?

Question 4: Iteration 1

Chart type: Bar Chart
Variables (encoding):
Avg Life expectancy (Y axis)
Avg Years of schooling (Y axis)
Yesr (X axis)
Regions (color)

PROS

- Show the trends of each region intuitively

CONS

- Hard to compare different regions
- Lack of details

DISCUSSION

- Maybe the line chart works better if the focus is growth rate.

Question 4: Iteration 2

Chart type: Line Chart
Variables (encoding):
Avg Life expectancy (Y axis)
Avg Years of schooling (Y axis)
Yesr (X axis)
Regions (color)

PROS

- Can compare the two variables of a specific region.

CONS

- These two variables are in different quantitative ranges. Life expectancy 0- 80, years of schooling 0-10. It might mislead users to put them on the same scale of y-axis.
- Hard to compare between countries and regions

Question 4: Iteration 3

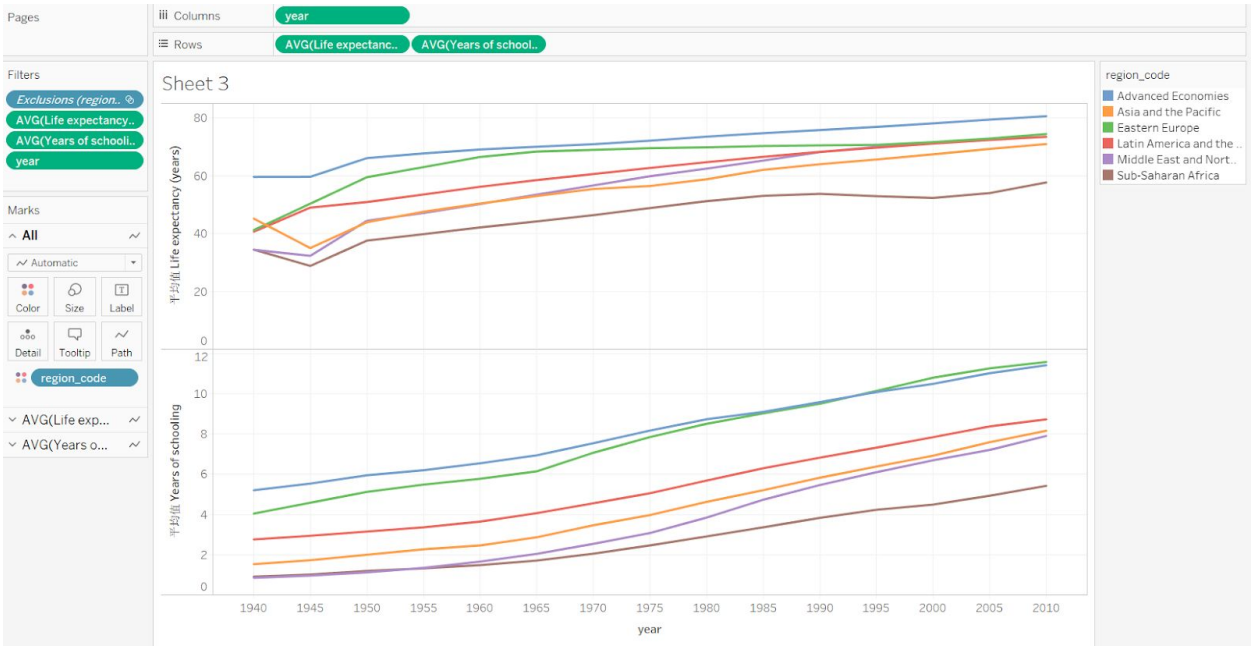


Chart type: Line Chart
Variables (encoding):
Avg Life expectancy (Y axis)
Avg Years of schooling (Y axis)
Yesr (X axis)
Regions (color)

PROS

- Show the trends intuitively
- Easy to compare between different countries
- Also indicate the rate of change

CONS

- Hard to compare these two variables together
- Lack of details

Initial Question 5: Do other attributes of countries have relationships with education level and life expectancy? Any interesting patterns?

Question 5: Iteration 1

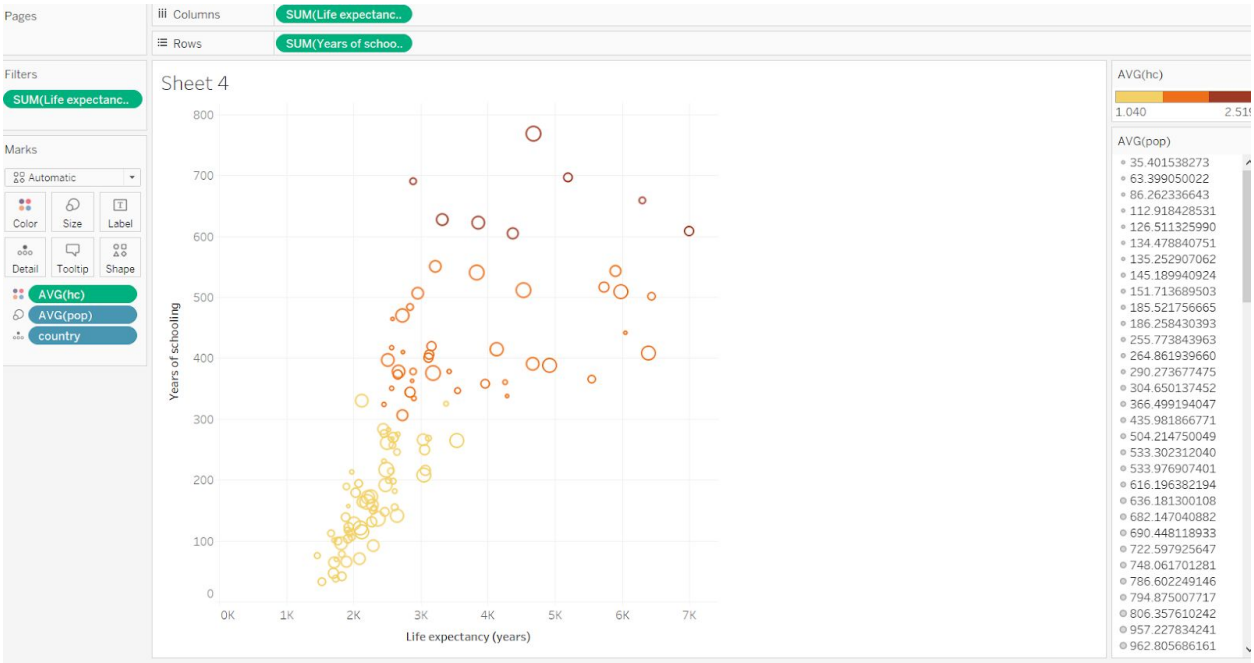


Chart type: Dot plot
Variables (encoding):
Avg Life expectancy (X axis)
Avg Years of schooling (Y axis)
Human Capital (Color)
Population (Size)
Country (Position)

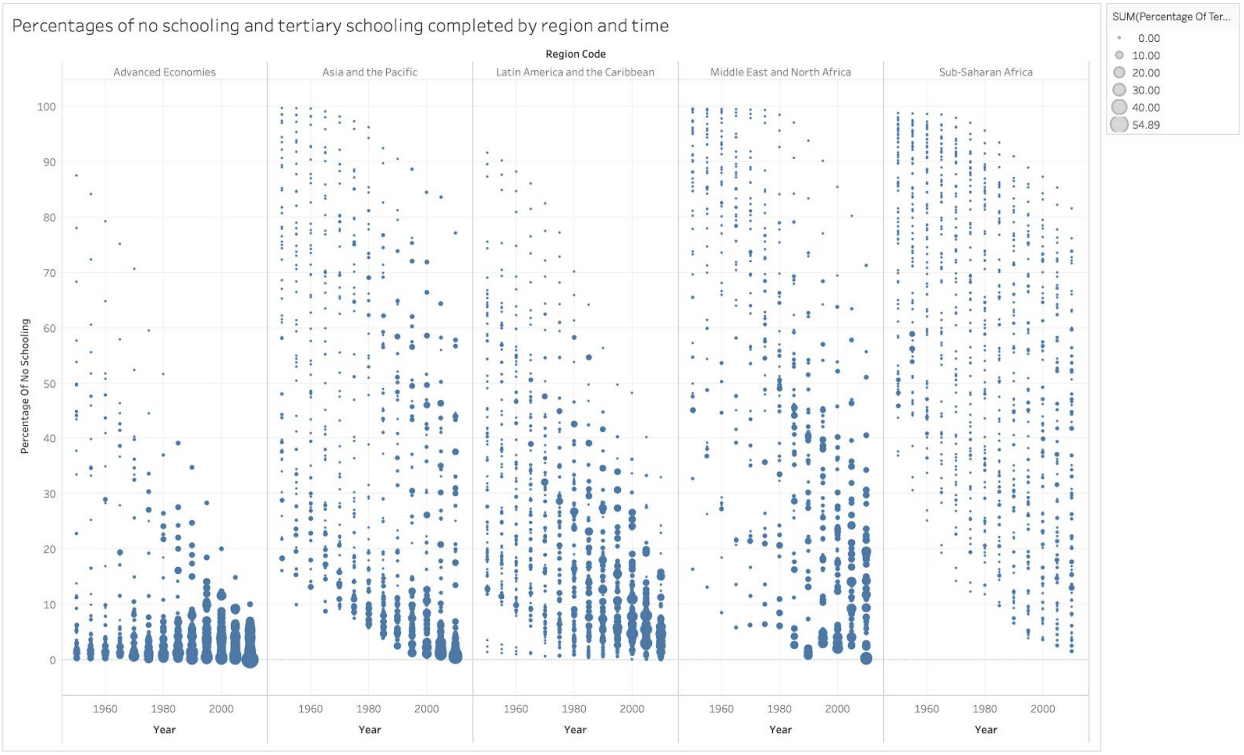
PROS

- It's easy to compare between different countries in terms of four quantitative variables

CONS

- Doesn't show the change with time

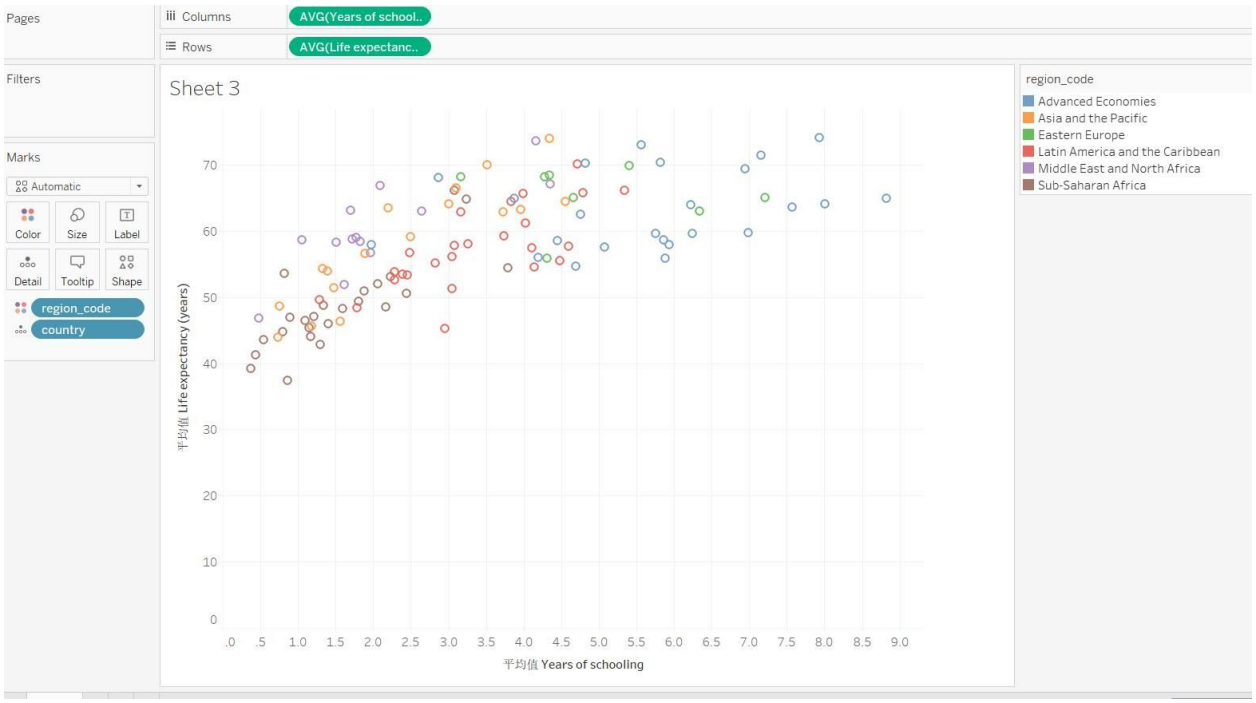
Final Visualization for Refined Question 1



Question 1: How do different regions (the Advanced Economies, Asia and the Pacific, Latin America and the Carribean, Middle East and North Africa, and Sub-Saharan Africa) compare in terms of changes in “years of no schooling” and “percentages of tertiary complete” from 1950 to 2010?

- It’s clear that the percentage of no schooling in all regions has moved from the higher end of the y axis to the lower end over time, suggesting a universal advancement in general education attainment.
- For regions including Advanced Economies, APAC, LATAM and Middle East and North Africa, there’s a trend of increase in tertiary education attainment with the decrease in no schooling. However, for Sub-Saharan Africa, the increase in tertiary education attainment over time is barely visible, even though it has experienced a similar downward movement in no schooling like the other regions.
- Advanced Economies and LATAM and the Caribbean have experienced the biggest drop in years of no schooling from 1950 to 2010 as evidenced by their steeper slopes compared to those of APAC, Middle East and North Africa, and Sub-Saharan Africa

Final Visualization for Refined Question 2



Question 2: Do different dimensions of education attainment correlate with life expectancy, and if so, in what regions?

When highlighting a specific region, it's easier to see the situation of this region.

- Advanced economies: Overall these countries have the highest population expectancy, without big differences. The years of schooling range from 2 years to 8.8 years. There is not a correlation between the two variables.

- Asia and the Pacific: There is an obvious trend that higher educational countries have higher life expectancy. Countries in this region don't have very high average years of schooling.

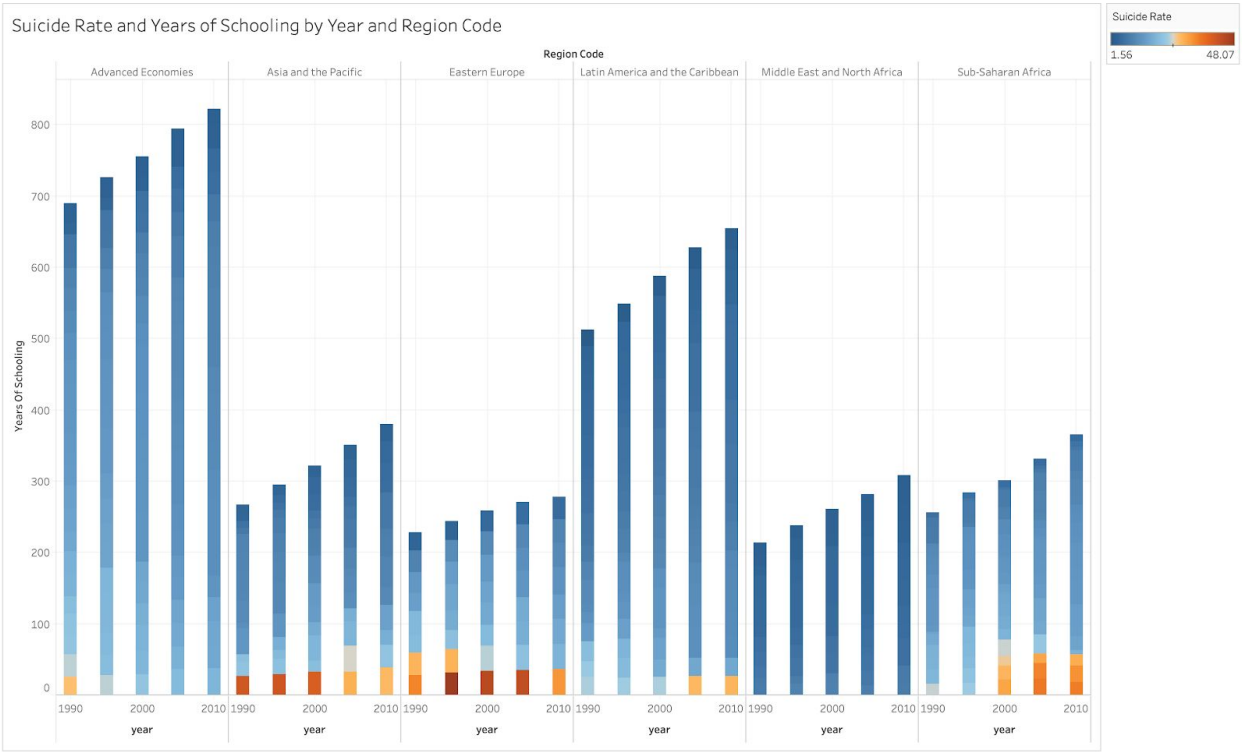
- Eastern Europe: Overall these countries have the highest population expectancy, with slight differences. No correlation between the two variables.

- Latin America and the Caribbean: There is an obvious trend that higher educational countries have higher life expectancy. Most countries are at the middle level in terms of educational attainment.

- Middle East and North Africa: There is an obvious trend that higher educational countries have higher life expectancy. Most countries in this area are relatively low-educated.

- Sub-Saharan Africa: There is an obvious trend that higher educational countries have higher life expectancy. Compared to other regions, both education level and life expectancy are the lowest.

Final Visualization for Refined Question 3



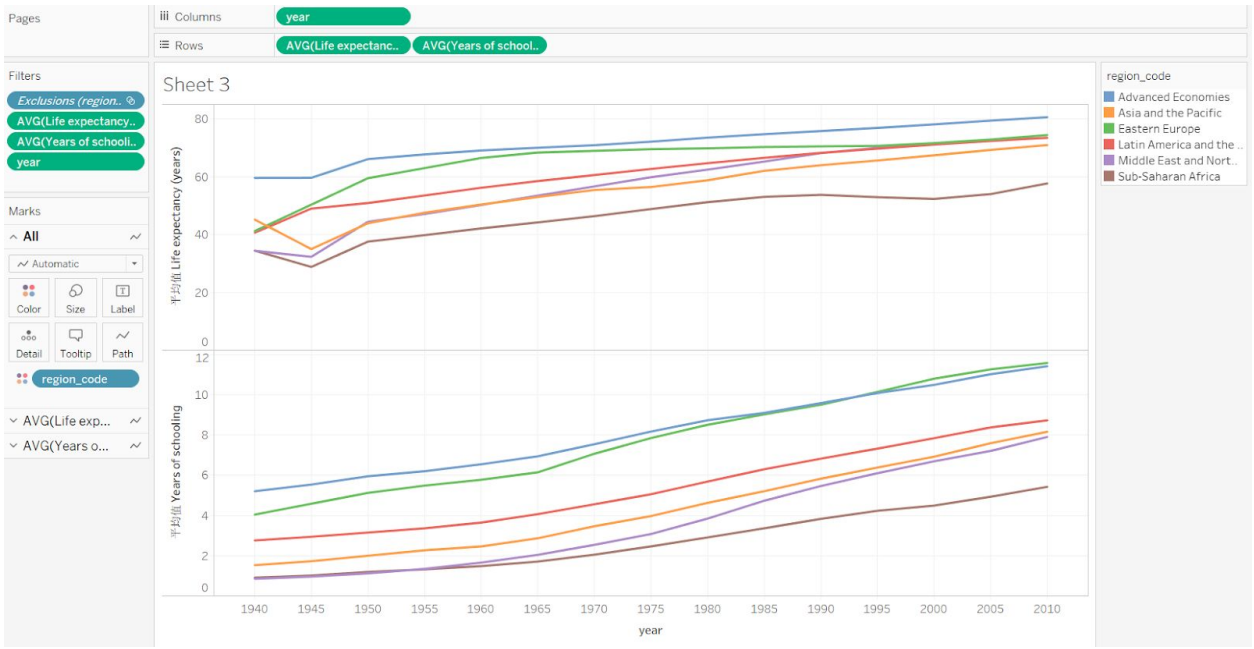
Question 3: How does average years of schooling correlate with suicide rate, and if so, in what regions?

- While all regions have experienced consistent growth in average years of schooling from 1990 to 2010, changes in suicide rate have not been consistent with the pattern of change in schooling and no clear trend can be identified for any region except the

Developed Economies.

- The suicide rate of Developed Economies has decreased with the increase in average years of schooling

Final Visualization for Refined Question 4



Question 4: Do different regions have different growth rates in educational attainment and life expectancy?

There are several regions undergoing a reduction in life expectancy from 1940 to 1945, might due to World War II. The growth of life expectancy in Estern Europe is relatively slow compared to other regions. Middle East and North Africa and Sub-saharan area start from the same level in 1940, in terms of education and life expectancy. However, Sub-saharan area develops much slower, which may be resulted from natural conditions.

Final Visualization for Refined Question 5

Question 5: Do other attributes of countries have relationships with education level and life expectancy? Any interesting patterns?

- In terms of human capital, it does show that countries with higher educational attainment and life expectancy come with higher human capital. the three step colors show the trend very clearly, with yellow dots staying in the left lower area and wine red in the upper right part.

- However in terms of population, it shows that some small countries (small populations) have very good performances in education, life expectancy and human capital,like Switzerland. While some countries with huge population have lower indexes in education and life expectancy.

Visualization Tools

We used Python to clean and combine data, and Tableau to explore and visually analyze our data.

Conclusion

Interesting trends have been identified regarding changes in different dimensions of education attainment for different regions across time. The assumed correlation between education attainment and suicide rate has not been consistenly found for all regions in all time periods. The tools provided to us by Tableau allowed for an intuitive way to play around with datasets, even though each view has its own limitations. In this learning process, we have been able to have a more visceral understanding of the advantages and drawbacks of different visual encoding methods and chart types. Going forward, we'd love to add more education dimensions into consideration when exploring their correlation with life expectancy and suicide rate, since the original education dataset offers rich layers of education data, which have not been thoroughly used or explored during this round. We'd also love to consider adding interactivity to our visualization by way of exploring data maps further, which isn't presented as a feature of the final artifact due to the constraints of the document. In addition, the aesthetics can be enhanced by fine-tuning the details of those charts to offer a more visually delightful user experience.