A Case Study: Visualization of Development Indicators in South East Asia

# Introduction

I chose to do a case study into development indicators as it was of great interest to me and with the rapid economic growth of many countries in South East Asia, it would be interesting to investigate their differences with the aid of visualization. All of the data I am using is from The World Bank, which is not only a very reputable source but also has a wide variety of indicators to use in analysis. This allows me to pose broad initial questions in a variety of fields and still have many more facets of data to delve into as I progress through my case study. I have chosen the six countries in South East Asia with the highest GDPs (as of 2017), as these are the countries that are likely to have experienced rapid economic growth in the past 60 years. This will then serve as the basis of my report as I investigate the many other development indicators of these countries such as health, education and pollution. The time series of the data means that I will be able to identify trends, with the comparison of information over time in an exploratory fashion being possible. This will allow me to highlight when particular countries entered a period of rapid economic growth, and whether other development indicators followed suit or fell behind.

# Data Fitness

The data comes in a form where the rows represent the varying fields of development indicators and the columns represent the years, starting from 1960 leading up to 2019. However, many of the fields do not start recording data until later years, depending on the country and when they started recording such indicators. This proves problematic when selecting fields and I must choose ones that have enough data to be visualized and with enough meaning to interpret. It may also be the case that some countries have recorded indicators that other have not or that they started recording much later – in which case I will select the indicators that I can use and disregard the others.

That being said the data does contain enough records in the fields needed to answer my initial questions, therefore making the dataset fit. It may come to light that more data is needed to answer my later questions during my exploration process, however I will comment on this during that section if required.

# Initial Questions

## RQ1: Which countries underwent rapid economic growth in the latter half of the 20th and early 21st Century?

This question is one of the most fundamental to the case study, as we need to understand if a country entered a period of rapid economic growth, and if so when this occurred. With this information we can then predict when the other development indicators of the country should have improved through a wealthier economy – or if not see how this compares to the other countries. Through this we should able to see which countries improved in many areas of their development indicators and which countries only developed in a few.

## RQ2: Did health increase throughout this period?

This question aims to give a broad idea of whether health improved across this period of time, with life expectancy at birth being a general indicator of this. In the modern era many developed countries have comparable life expectancies, but this question will also shed light on which countries potentially had poorer life expectancies and improved greatly with the rapid economic growth of the country.

## RQ3: Did education improve during this period?

Another key development indicator for countries is education, which is often used alongside wealth and health to determine a country’s development. The number of pupils in secondary education again is a general metric but should give a broad idea of whether education in the general population increased with a growth in the economy.

## RQ4: Did pollution increase during this period?

In many first world countries a growth in their economy is linked to a growth in industry, which in turn also gives rise to the amount of pollution produced. With countries going through rapid economic growth, this question aims to show which countries produced the most air pollution during that time.

# Basic Visualization

First the data is loaded into R as a data frame containing all the indicators for all the countries. Next it is subdivided into individual data frames for each country, making it easier to logically subset and produce visualizations. I created a function called “subsetSingular” which was used to clean and transform the data into a usable format for visualization - taking a country’s data, the series name and year selection as parameters. The functions filter and select are used to first subset the required series and then the years, which is needed to clean the data to remove any N/A values. The data is initially in the form of using rows for the series and columns for the years, meaning that the data needed to be transformed. In the first question it made the most sense to use the years for the x-axis variable, which meant that the row names would need to become a column of data in themselves. Due to the way in which R stored the data in the data frame, various changes of types and column names were necessary to transform the data into a usable format.

For each of the initial questions I decided to visualize a graph for each country. This allows for easy comparison of not only trends but also for absolute values, allowing the viewer to gain the most information from the data.

## RQ1

The series I chose to visualize was each country’s GDP (constant 2010 US$) against each year. I chose this because if a country had undergone rapid economic development it would be most apparent in an increasing GDP over time. The visual encoding I chose was a line graph because both the year and GDP are quantitative data, meaning that position would be the most effective means of encoding. This visualization would technically be classed as univariate as only one variable (the GDP) is being measured on the y-axis and the year is the data case along the x-axis. I also thought that a line graph would be most appropriate because I am using time series data, of which line graphs are very effective at showing trends.

A close up of a map

Description automatically generatedFig. 1 Six line graphs to show the increase in GDP (constant 2010 US$) from 1960 to 2019 for each country

## RQ2

As previously mentioned, life expectancy at birth was chosen to represent health as it gives broad estimation of the improvements made to a country’s health system. I wanted to compare the life expectancy to the GDP of each country making it a bivariate data plot, and so I chose a scatter plot with a line of best fit. A scatter plot only uses position to encode the data making it the most effective means, with the added colour of the trend line to distinguish it from the points. Both variables are quantitative with the aim to show the relationship between the two, making a scatter plot the most appropriate encoding. The life expectancy was the variable being measured and so this was placed on the y-axis and the GDP on the x-axis. Since each x and y coordinate still correlates to a year, the graph would still show the relationship over time, meaning it could be compared to the first plot used in RQ1.

A close up of a map

Description automatically generatedFig. 2 Six line graphs to show the relationship between GDP (constant 2010 US$) and life expectancy at birth, total (years) for each country

## RQ3

Similar to RQ2, I wanted to show the relationship between two quantitative pieces of data over the given period of time. I chose to measure the number of secondary school pupils as an indicator of education, with the premise that a country with an increasing level of education would have an increasing number of pupils in secondary education. Again, I wanted to plot this against GDP to see if an increase in the GDP would have a positive effect on the education of a country. This meant that the same visual encoding would be fit for the same reasons, with the line of best fit through the scatter plot showing a trend.

Unfortunately, Japan had no records with the World Bank for this data series, however it was still the series with the most data for all the other countries I was investigating, so continued to use it and omit Japan.

*A close up of a map

Description automatically generated*Fig. 3 Six line graphs to show the relationship between GDP (constant 2010 US$) and the number of pupils in secondary education for each country (Japan omitted due to lack of data)

## RQ4

This question was also similar to RQ3 and RQ2, with two quantitative pieces of data over a given period of time, meaning a scatter plot would be the most appropriate to show the relationship. This time total greenhouse gas emission was being measured so that was placed on the y-axis and again it was being compared to GDP which was placed on the x-axis. In this question the scatter plot seemed to be the most beneficial with much more erroneous data being present. Where in the previous questions a line graph could have potentially been used for the time series nature, it would definitely not have been an optimal choice for this visualization.

As you can see the scale of the x-axis varies between each country, due to some countries not producing greenhouse gas emission data for some years, meaning that some of the graphs are not representative of the whole time period. This was far from ideal but in all cases, there was enough data to show a trend. In the case of Indonesia, the graph began to show a positive trend but then dropped to negative before they stopped producing data for this series. The numbers from Indonesia vary wildly and I am inclined to treat this country as an anomaly, taking the positive correlation from other countries to be the general trend.

A close up of a map

Description automatically generatedFig. 4 Six scatter plots, with lines of best fit, to show the relationship between GDP (constant 2010 US$) and total greenhouse emissions (kt of CO2 equivalent) for each country

# Exploratory Process and New Question Generation

For my exploratory process I wanted to expand on each of my initial questions, with at least one further question in that area of the case study. I set out with the aim of looking at the visualizations from each of the initial questions and thinking of new questions that they gave rise to. This would then allow me to delve deeper into that topic and use more specific indicators from the data to answer these questions. This worked well and I was never stuck on finding new questions from the data. However, I had predicted earlier that many indicators would not have enough data to produce adequate visualizations and I was correct, meaning that I had to choose questions where enough data was present. This for the most part worked well, and I was able to adapt certain visualizations where there was little data to still answer the question satisfactorily.

## FQ1: How does the overall economic growth percentage of each country compare?

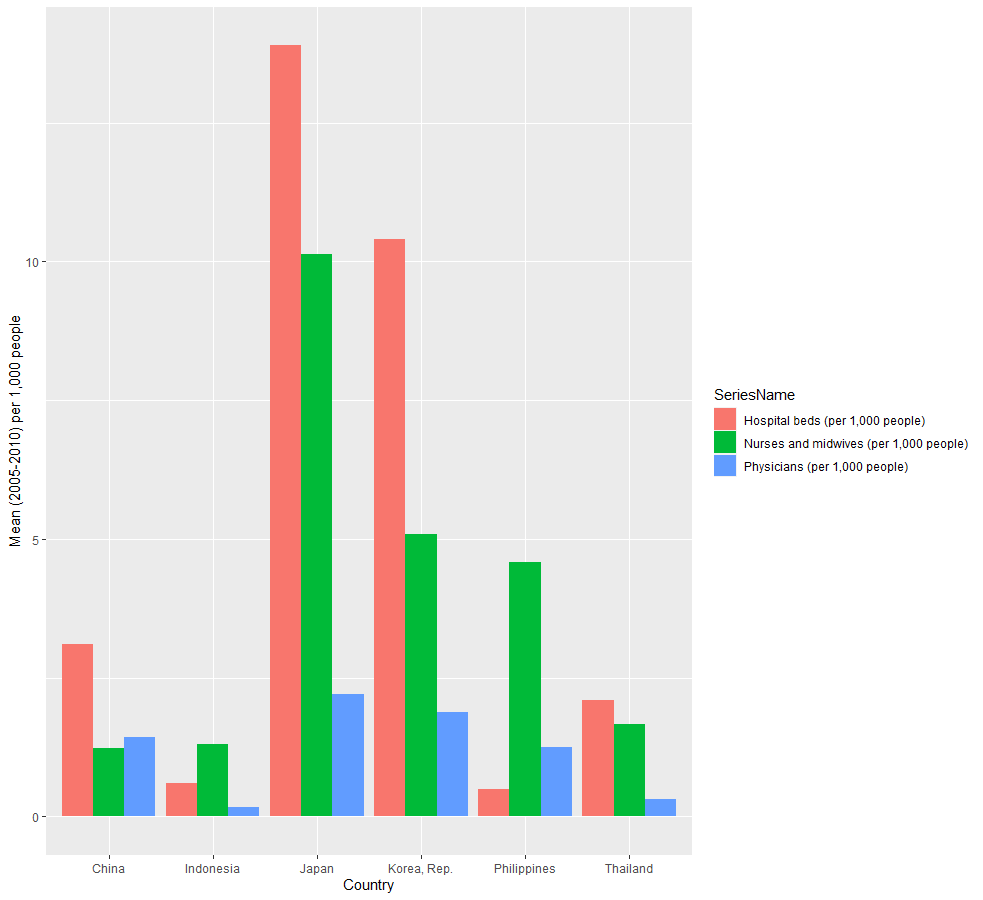
For the GDP of each country the first visualization showed that every country had a large increase over the time period, but not which country had increased its GDP the most as a percentage of itself. Therefore, this is the question I wanted to answer in this topic, with a visualization of the sum of each country’s GDP growth percentage.

A picture containing screenshot, table, drawing

Description automatically generatedFig. 5 A bar chart to show the sum of GDP growth (annual %) for each country from 1961-2018

## FQ2: Why do some countries have higher life expectancies than others?

The initial question on the topic of health showed the general trend of life expectancy increasing with a country’s GDP. But the data showed that there was a large variation between countries in regard to their most recent life expectancies, which led me to this question. The data provided three useful indicators to look into this – hospital beds; nurses and midwives and physicians (all per 1,000 people). The visualization would then hopefully show which of these indicators had the largest difference and was correlated with a higher life expectancy. Unfortunately, not all of these metrics were reported frequently, so I decided to take an average from 2005-2010, where all countries had a good amount of data within these years.

Fig. 6 A grouped bar chart to show the number of hospital beds, nurses and midwives and physicians per 1,000 people from 2005-2010

## FQ3: In China and South Korea, was the drop in secondary education pupils a result of a lower population in the secondary education age range?

On the topic of education, the visualization of the initial question showed a positive trend between the number of secondary education pupils and GDP for three out of five countries. In China and South Korea, the numbers began to drop in the 21st Century despite their GDPs still increasing. This was against what I had predicted the trend to be and decided to investigate why this was. Because the indicator was based on the number of people between a certain age range, I decided to investigate this over the time period, to see if there was a decline in the number of students because of a decreased early age population.

A close up of a map

Description automatically generatedFig. 7 Two line graphs to show the population between the ages of 0-14 from 1960-2018 for China and South Korea

## FQ4: Which sectors were the biggest contributors to CO2 emissions in China?

In the scatter plot from RQ4, China had a consistent increase in their overall air pollution and produced the most out of the six countries. This led me to investigate which sectors in China were producing the most pollution. The initial indicator accounted for all greenhouse gas emissions; however, the data did not have many detailed indicators for all greenhouse gases and CO2 appeared to be the standard for measuring each sectors contribution. The data had broken down the country’s total CO2 production into five sectors as a percentage of the total – Electricity and Heat; Manufacturing Industries and Construction; Transport; Residential Building and Commercial and Public Services and Other Sectors. I decided to visualize three different years, with jumps of 20 years to appropriately show that changes that occurred over time. See the Appendix for the full visualization.

# Visualization Design

In FQ1 I chose a bar chart because I knew the countries were going to be used as the data cases (along the x-axis) and the dependent variable was the sum of the GDPs (on the y-axis). In this case the countries are nominal data meaning that position, as with all data types, was the best to visualize the differences effectively. The sum of the GDPs is quantitative data, but since position was best suited to distinguish between countries, the next best visualization was length – used in the length of each bar. The use of a bar chart makes it easy to compare univariate data such as in this case, clearly distinguishing if one variable is larger or smaller than another. This is one of the main reasons I didn’t choose a pie chart, where these small discrepancies could be missed.

For FQ2 I knew that I was going to be working with tri-variate data so I would need to make sure I represented each variable in its own explicit way. Similar to FQ1 I needed to compare the quantitative data easily, therefore I opted for a bar chart again. This worked well because only one of the variables was quantitative, with the other two being nominal (the country and series). I used the x-axis to distinguish the countries and hue for the series, as it is the second most effective at distinguishing nominal data. I used a grouped bar chart, where each of the of the series was a different colour in the same country group, producing an effective visualization. I opted for a grouped bar chart over, for example, a stacked one because it would be easier to compare the individual series with all of them based at zero.

The visualization for FQ3 turned out to be very similar to RQ1, but for two countries instead of all six. The question needed the visualization to show a trend over time, for which I used a line graph. Both the year and population are quantitative data, but with the population being the dependent variable I placed it on the y-axis and the year on the x-axis. Using a line graph allowed me to use position to represent both variables, making it the most effective visualization possible.

None of my visualizations had yet been interactive and for FQ4 I saw this as the perfect opportunity to do so. I used R to subset the data as I was already familiar with how do so easily and opted to use D3 to create an interactive visualization. Given the data I had initially thought to create a pie chart to show the percentages, as it would better visualize each sector as out of a whole 100%. Each sector was nominal with a quantitative value, making it suitable, with each year producing an individual chart. After some research I changed the pie chart to a donut chart as it proved to be a more beneficial visualization to the user, taking away from comparing the slice areas and focusing on the arc lengths. This improves differentiating similar sectors and also makes a more efficient use of the space, if I were to place any information in centre. Additionally, I added a feature to show the individual values of each sector when the mouse is hovered over them, to further distinguish the differences. The labels with lines and colour coding scheme are used to clearly identify each sector, with each sector being kept the same colour and in roughly the same position for each chart. The buttons at the top of screen allow the user to switch between years, with an animation being used to help accentuate the growing or shrinking of each sector, while also helping to engage the user more.

# Reflection

Upon reflection of my development process I am happy with the outcome. I had potentially planned to do another interactive visualization similar to that of Hans Rosling from the TED talk but for my data, however I had limited knowledge of D3 and it was not possible. The majority of my visualizations were time-orientated, which meant I couldn’t produce some types of visualizations, such as those centred around distribution like histograms and box plots. Although I believe the ones I did produce were insightful into the data and produced information that wouldn’t be known by just looking at the data, with the exploratory process helping me ask questions I wouldn’t have otherwise asked.

# Appendix

## FQ4 Screenshots for each year of the interactive visualization

*A picture containing device

Description automatically generated*Fig. 8.1 A donut chart to show the proportions of CO2 production from each sector in China in 1971

*A picture containing device

Description automatically generated*Fig. 8.2 A donut chart to show the proportions of CO2 production from each sector in China in 1991

Fig. 8.3 A donut chart to show the proportions of CO2 production from each sector in China in 2011

*A picture containing device

Description automatically generated*