Guiding Question

About 50 years ago, when China just finished its four-year civil war and had its new regime established, every industry was waiting for recovery. The development of the society needs manpower and under this consideration, new China's government encouraged its people to bear as many children as they can. Both of my parents were born in those years. At that time, it was very common that a family has 4 or 5 children. Mother who bore lots of kids was known as "Mother Heroine". This movement significantly increased population of China and ended up with another well-known movement - "the One Child Policy".

What interests me is that people in old China used to be proud to have a big family while as is well-known, a lot of developed countries are facing the problem of negative birth rate and population decrease nowadays. Similar issue exists in China's new generation as well. With more and more people have the opportunity to receive higher education, it seems people are less willing to establish family or have kids. Is this the effect of education? Is that the more education people get, the less they are willing to have children?

Therefore I decided to study the **effect of education on people's willingness to have children**.

To prove my guess on a data-driven way, I decided to find data that can gauge:
1) People's chance to receive education of a country
2) People's willingness to have children of a country
And then compare across different countries to get a conclusion.

For 1), I found the following data sources:

a. GDP per capita
   The higher GDP per capita a country has, the wealthier the country is, and people would have higher chance to receive education.
b. Total public spending on education as percentage of GDP
   The higher public spending on education is the higher chance people would have to receive education.

For 2) I found the following data sources:

a) Birth rate
b) Wanted Fertility Rate (births per woman)
c) Population growth rate

For countries, I want to have as plenty countries as possible and I also want to compare across developed, developing and underdeveloped countries.

My response

After deciding the guiding question and data sources I need, I started to find the corresponding data in World Bank's database via the *wbdata* module in python.

One major problem I found is the **data is not well populated**. Usually the indicator exists and I can find download a list of data. However the value shows "None". Therefore, the first thing I did is to write **a function that can tell if the data is available,** given country and indicator.

The second problem needs solving is even if I can assure the data is available by using the function I mentioned above, the **data is not ready to use immediately**. The data is contained in a long dictionary. Therefore, the second thing I did is to write **a function that can extract data** from the dictionary and return a list of ready-to-use data.

The third problem is the order of data I got from the function I used above is different from the one I put in the argument. Therefore, to make sure the index is consistent with data, I wrote **a function that extract list of country names** that have the same order with data generated above.

When thinking of the problem, I found it's also interesting the effect across income levels and region. To get these data, I used information from *search_countries* function in *wbdata* module.

To wrap up, I put everything into panda data frame and upload to MySQL, which can be read back in R using RMySQL package. Till here, all data preparation work is finished.

What I wanted to do and what I ended up doing

As described in my guiding question, I wanted to compare the effect of education on people's willingness of having children, especially the similarity or difference between China and western developed countries. However, it ended up that the *public spending on education* data is not available for most of the years for China. To assure the result is most updated, I ended up giving up China and turned to research a more generalized question.

Another thing I ended up doing is I chose fertility rate instead of wanted fertility rate as my research variable. This is because the data wanted fertility rate is not available for most of the countries in database. I could have chosen the countries which have wanted fertility rate data available. However, most of the countries available turn out to be developed, high-income, European countries. This renders my study much less reliable. Also, there's obvious overlap between birth rate, population growth rate and fertility rate. Therefore, I decided to use fertility rate alone, which has most countries data available, as my research variable.

Workflow

Part I: Data Collection and Transformation

Step 0: Test if data for certain indicator, country and year is available.

Use function *test_year()*. Here's an example:

```
>>>test_year("SP.DYN.WFRT", “ZMB”, 2005)
Country: Zambia
Data: Wanted fertility rate (births per woman)
Year: 2005
Staus: NOT available
```

Step 1: Retrieve data from API and store in panda data frame.

1) Make sure all wanted data is available by repeating step 0
2) Note that data retrieved with wbdata module are in “string” type. Conversion to “float” type is needed if data is numeric.

Use function *get_value(), get_countryList(), get_region()* and *get_incomeLevel()* to retrieve data and use function *float()* to convert data type is data is numeric.

a. Indicator for wanted data:
   GDP per Capita: NY.GDP.PCAP.KD
   Expenditure on Education: SE.XPD.TOTL.GD.ZS
   Fertility Rate: SP.DYN.TFRT.IN
b. 33 countries are studied. Label for each country can be searched by function *wbdata.search_countries()* .
c. Function *get_value()* and *get_countryList()* both take three arguments: *indictor*, *country* and *year*. Argument *indicator* and *country* can be either a string or a list of string. Argument *year* should be an integer, since we only study the effect within a time frame of the same year.

Step 2: Upload data to MySQL database

Database needs to be created before uploading data. (Use MySQLdb module to create new database or delete existed database in MySQL)

Part II: Data Analysis

Step 3: Read data from MySQL in R

Step 4: Plot graphs with “*ggplot2*” package.

1) *ggplot* bubble chart for Expenditure on Education, Fertility Rate and GDP per Capita. Size of bubble stands for amount of GDP per Capita
2) *ggplot* bar chart for Fertility Rate and Region.
3) *ggplot* bar chart for Fertility Rate and Income Level

Distribution of Tasks

Python - Data collection and data transformation

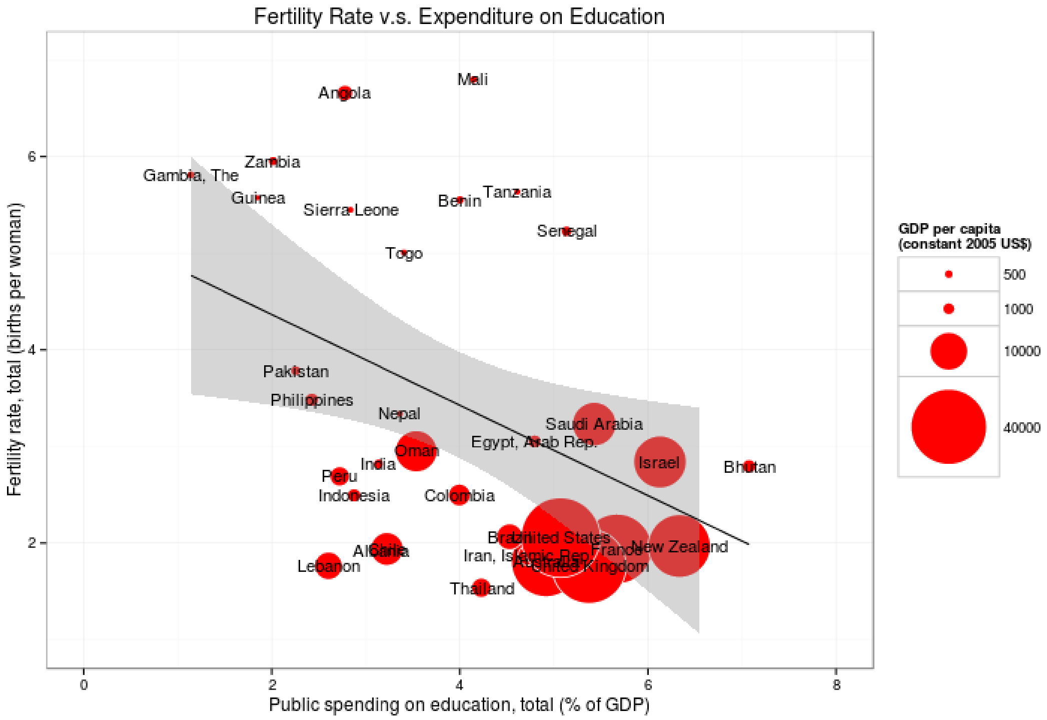
1) Retrieve data using *wbdata* module and other self-defined functions
2) Store data into data frame using *pandas* module
3) Creating and deleting database in MySQL using *MySQLdb* module
4) Upload data frame to MySQL using *pandas* module

R - Data Analysis:

1) Download data frame from MySQL using *RMySQL* package
2) Plot graphs using *ggplot2* package

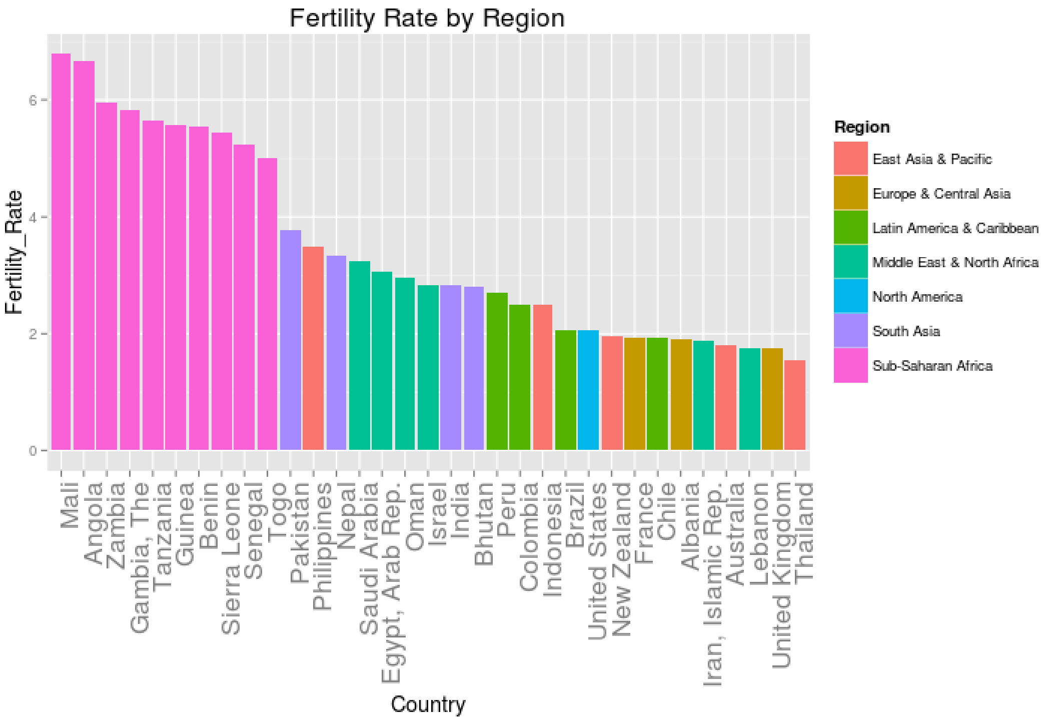
Show results using graphs and plots

1. Bubble chart for fertility rate and expenditure on education



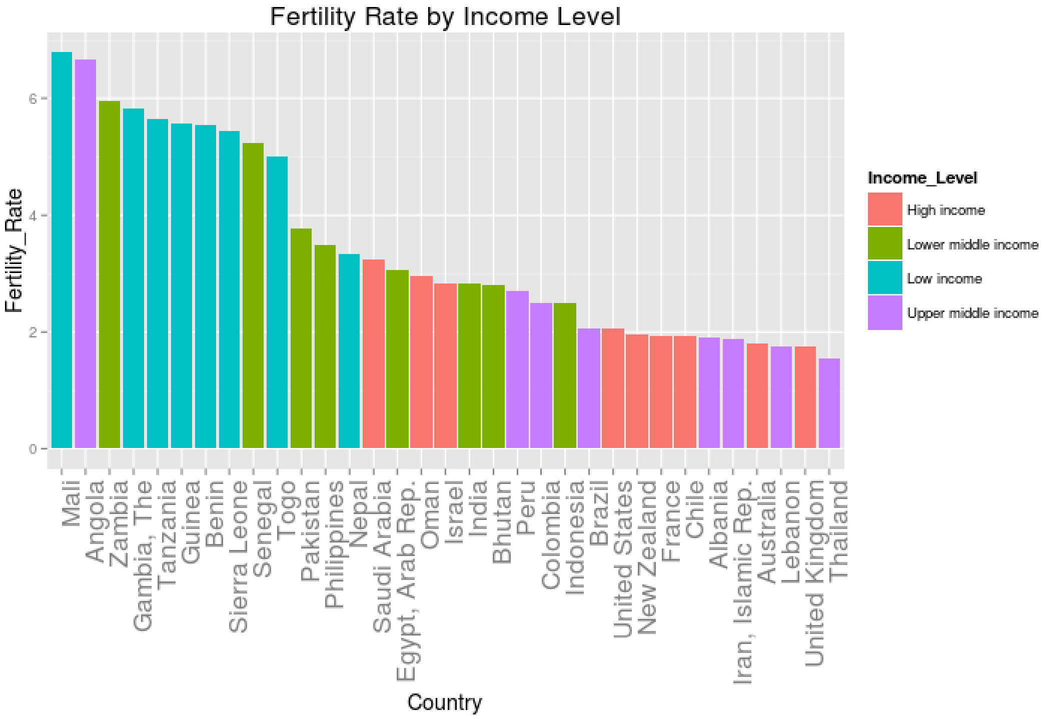
This graph shows that there is correlation between public spending on education with fertility rate, interpreted as the higher public spending on education, the lower fertility rate of the country is. In the meantime, it can also be observed that big circles tend to sit on the lower part of the graph, while small circles tend to sit on the upper part of the graph, which indicates that the higher GDP per capita of a country is, the lower fertility rate is. This can also be interpreted as the higher GDP per capita of a country is, the more likely the country's public spending on education to be high. The shadow are represents the 95% confidence level of the trend line. Therefore, we can conclude that education has effect on people's willingness of having children in a country.

2. Bar chart for fertility rate and region:



We can see that the top fertility rate countries are mainly from Africa.

3. Bar chart for fertility rate and income level:



This graph shows similar information with last one. It shows that most top fertility rate countries are low income countries.