Codecademy Assignment

Data Visualization with Python

A Codecademy Course Project

Life Expectancy at Birth and GDP: A Complicated Relationship

Background on the Project

As a capstone project for the Codecademy course on Data Visualization with Python, I am tasked with looking at the relationship between Gross Domestic Product (GDP) and Life Expectancy at Birth in Years for 6 Countries: Chile, China, Germany, Mexico, the United States of America, and Zimbabwe over the years 2000-2015.

What is Gross Domestic Product (GDP) and how is it calculated?

Gross Domestic Product is a measure of the monetary value of goods and services that a country produces in a specific year. Gross Domestic Product = private consumption + gross investment + government investment + government spending + (exports-imports). Countries can estimate this figure in different ways.

What is Life Expectancy at Birth in Years (LEABY) and how is it calculated?

Life Expectancy at Birth is defined as the average number of years that a newborn is expected to life given current mortality rates. Life expectancy at birth is an output of life table, a set of tabulations describing the liklihood of dying, death rate, and nubmer of survivors for each age group. The WHO creates these life tables based on about 1800 life tables based on "good quality" vital registration records. Where countries have vital

based on about 1800 life tables based on "good quality" vital registration records. Where countries have vital registration, the WHO assesses the level of completeness of this data and adjusts the mortality rates according to this assessment. In cases where there are not adequete sources of age-specific mortality rates, the life table is derived from estimated under-5 mortality rates and adult mortality rates applied to a global standard (the average of all 1800 life tables) using a modified logit model.

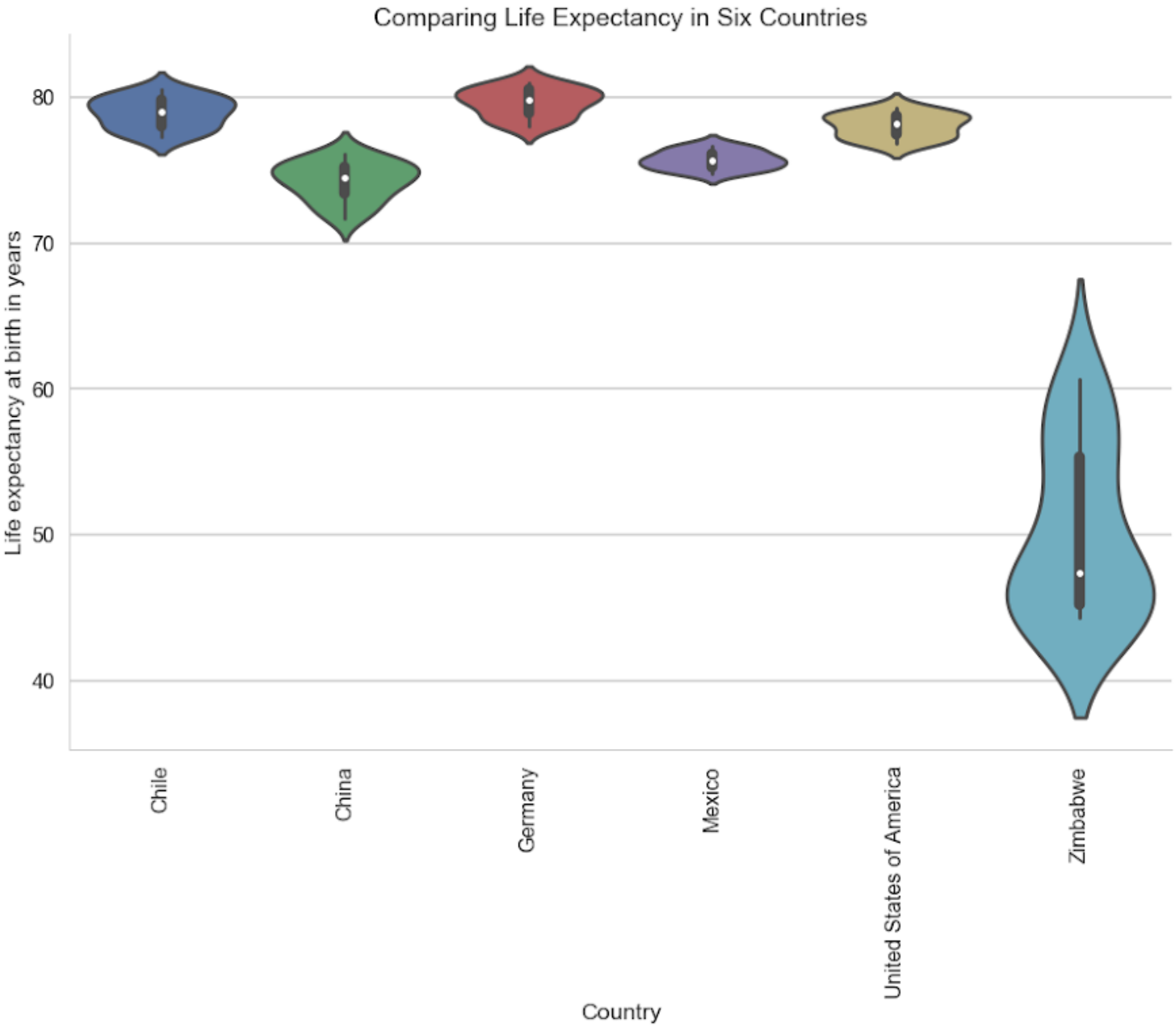
According to the WHO "The lack of complete and reliable mortality data, especially for low income countries and particularly on mortality among adults and the elderly, necessitates the application of modeling (based on data from other populations) to estimate life expectancy."

What is the Rural Population as % of Total Population and how is it calculated?

According to the UN Population Division, Total population is defined as "De facto population in country, area or region as of 1 July of the year indicated, while rural population is defined as "de facto population living in areas classified as rural according to the criteria used by each area or country".

Data Sources

GDP Data for this project comes from the World Bank National Accounts Data and the Organization for Economic Co-operation and Development (OECD). Data on the rural population as a percentage of the total population also comes from the World Bank's Data Portal. These are estimates based on the United Nations Population Division's World Urbanization Prospects: 2018 Revision. Data on the Life Expectancy at Birth in Years for each country was obtained from the World Health Organization's Global Health Observatory.



Comparing Life Expectancy at Birth in Six Countries 2000-2015

Violin plots are similar to box plots, but also overlay a kernel density estimation to provide information on the distribution of the data. In this violin plot, the width of the plots have been standardized.

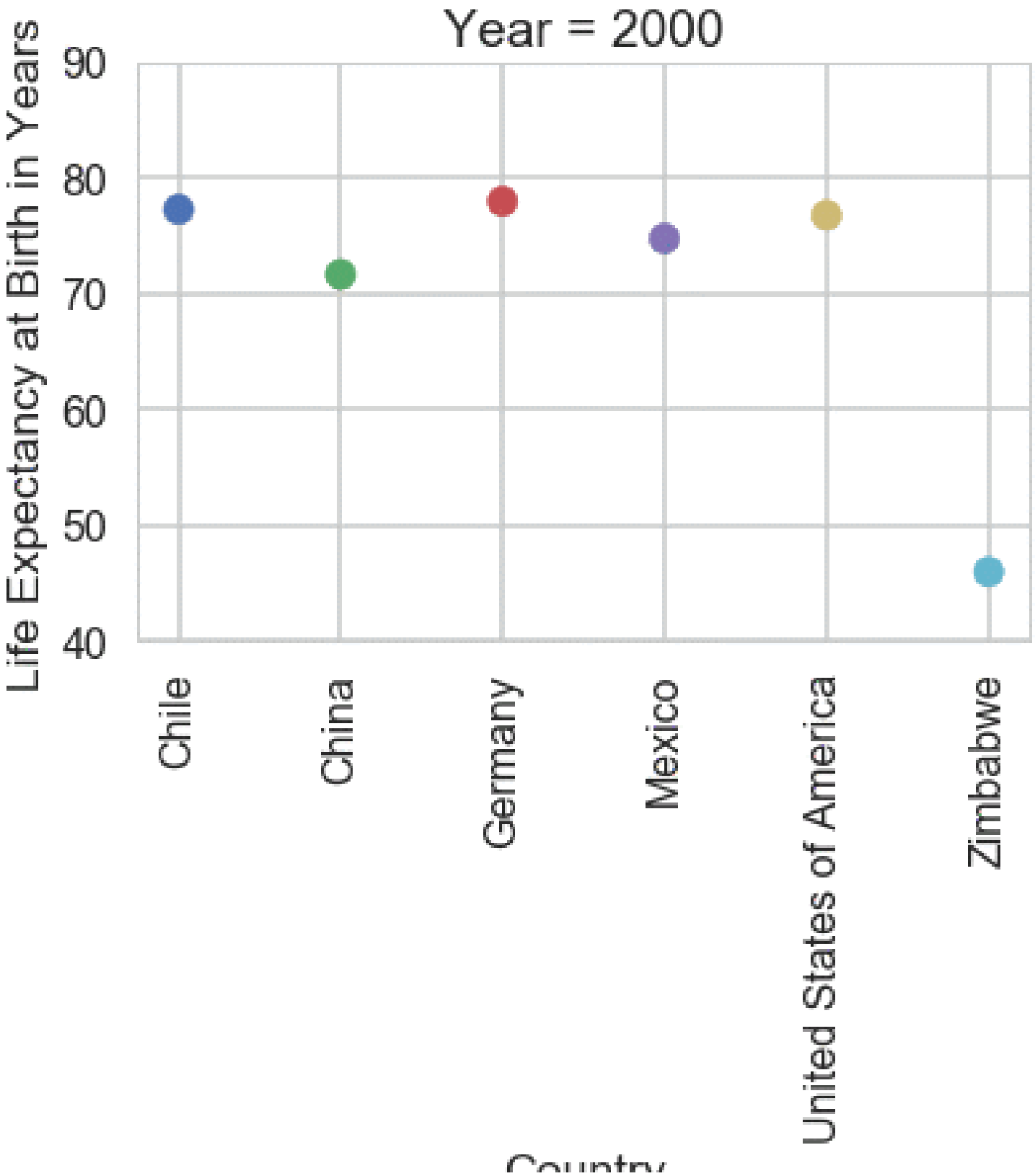
The mean of each data set is marked by the white dot. As you can see, Germany and Chile have the highest mean life expectancy, while China and Zimbabwe have the lowest mean life expectancy in years.

Zimbabwe and China also have the most variation in life expectancy over the 15 years of data under observation. Mexico has the least variation over this time period.

An animation as an alternate way to view the same data

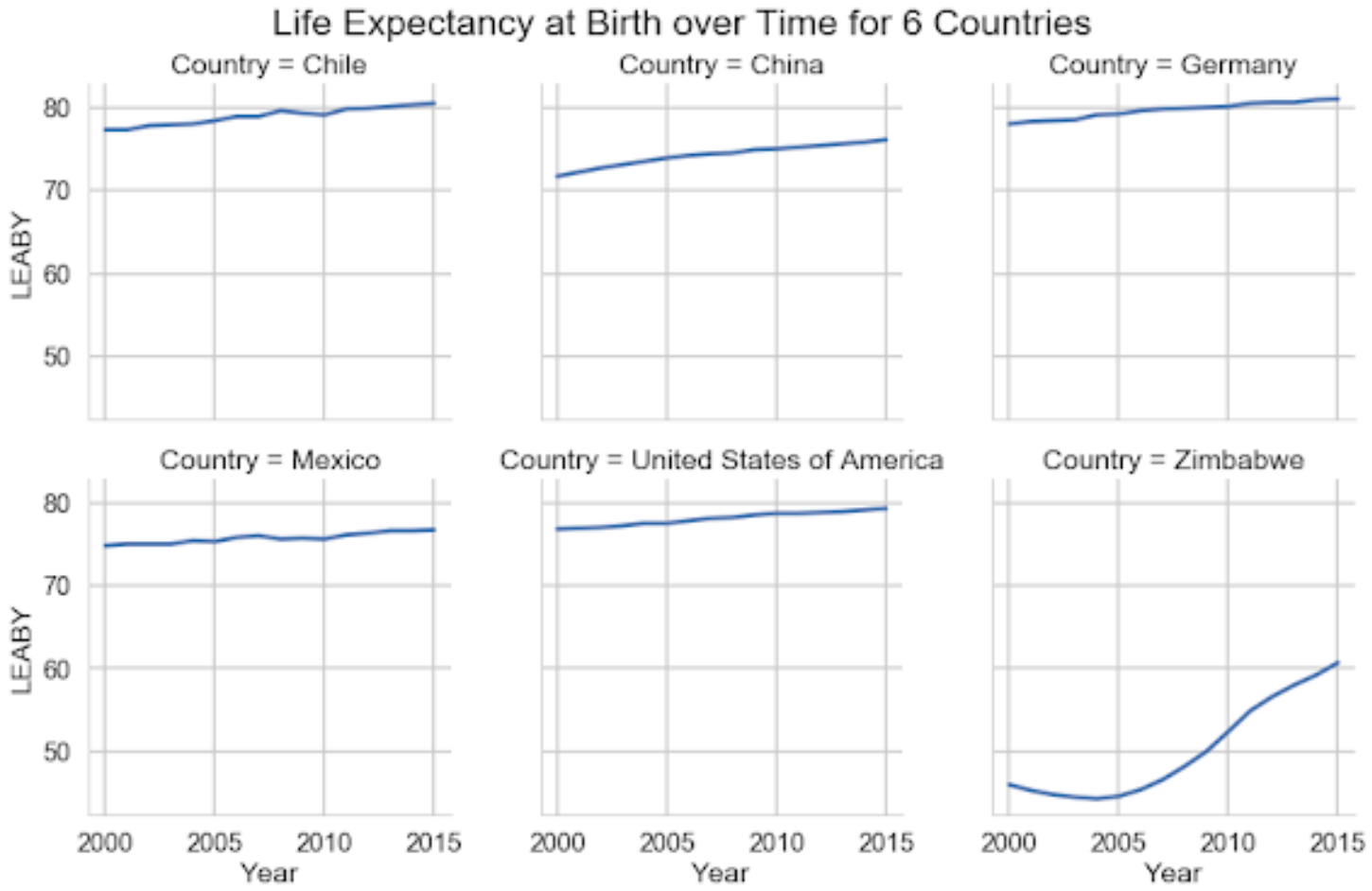
Another way to show how life expectancy changes over time is through an animated .gif that cycles through data from each year.

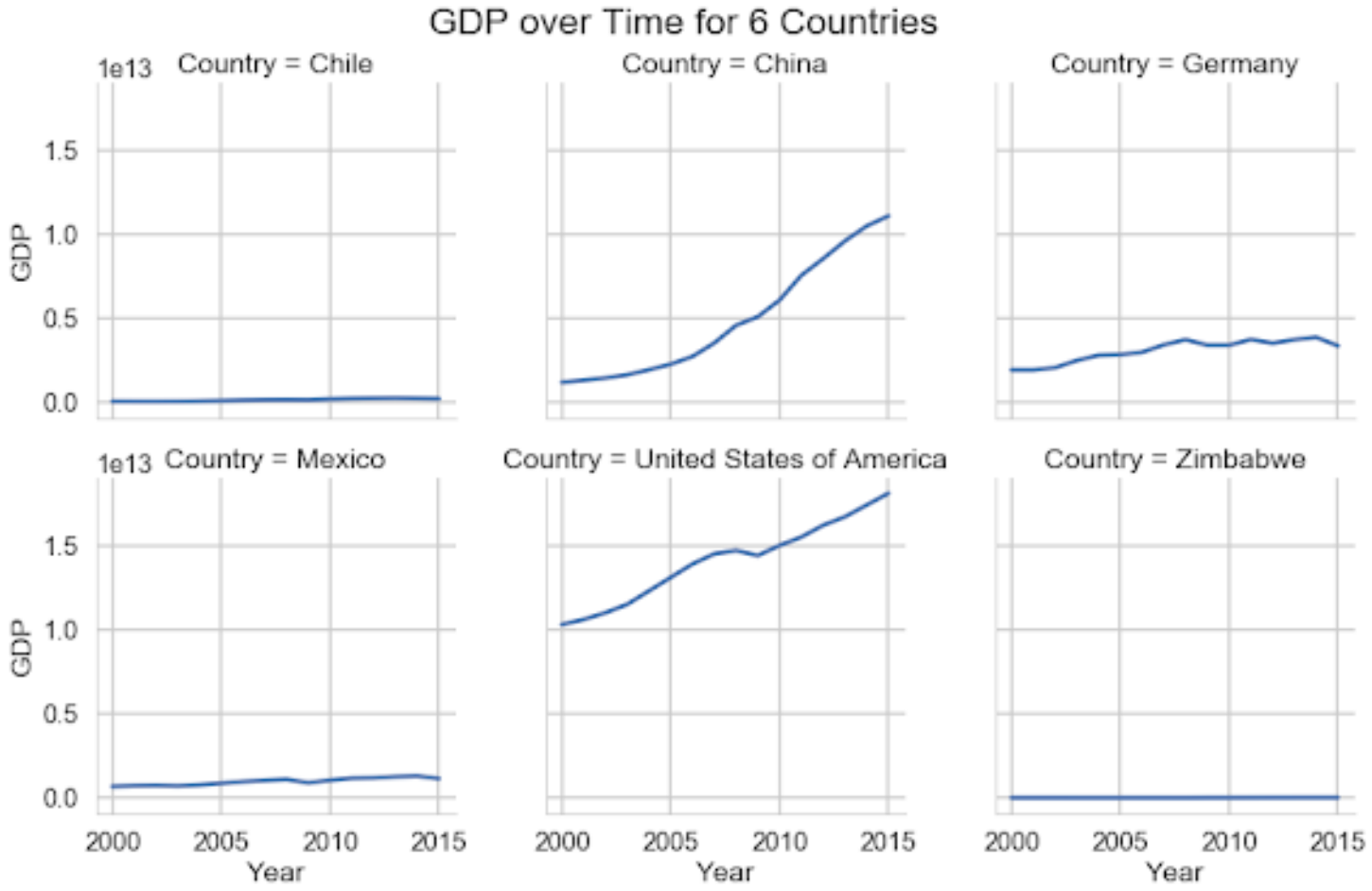
I spent some time trying to plot both GDP and LEABY on an animated plot and ran into a lot of trouble in including a legend that would plot colors to countries, so unfortunately was not able to include that figure here (I will continue to work on getting that code to work, though, since I have some ideas of what might be going wrong; I think the data is considered a series and as such I can't include separate legend entries for each point unless I somehow plot each point separately).



Country

Are Life Expectancy at Birth in Years (LEABY) and Gross Domestic Product (GDP) Correlated?


Life Expectancy at Birth over Time for 6 Countries
Country = Chile
Country = China
Country = Germany
Country = Mexico
Country = United States of America
Country = Zimbabwe
LEABY
80
70
60
50
2000
2005
2010
2015
Year


GDP over Time for 6 Countries
1e13
Country = Chile
Country = China
Country = Germany
Country = Mexico
Country = United States of America
Country = Zimbabwe
GDP
1.5
1.0
0.5
0.0
2000
2005
2010
2015
Year

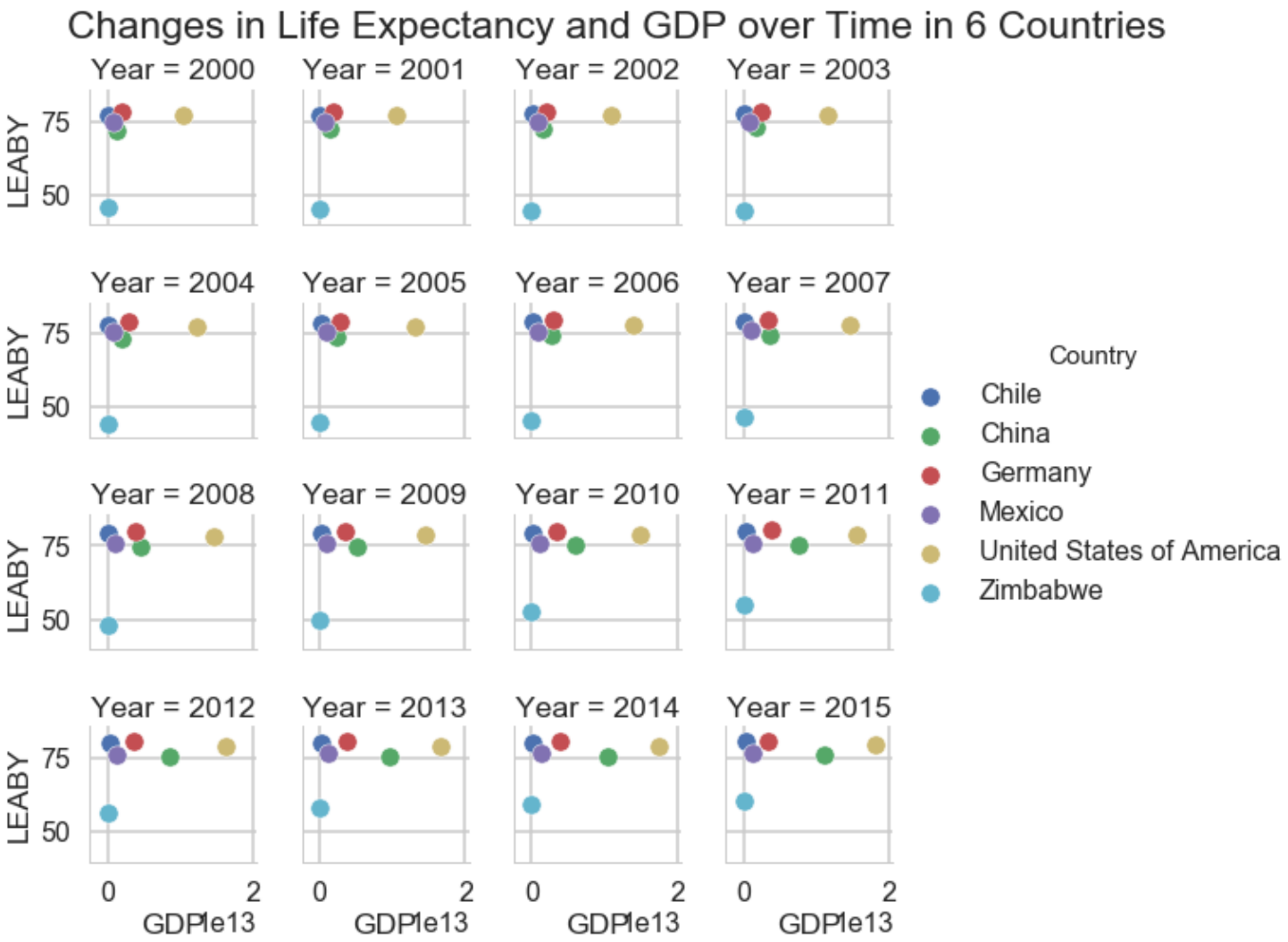
In the two charts above you can see more information about life expectancy (LEABY) and Gross Domestic Product (GDP) for each of the six countries.

Another way to look at the same data is displayed in the plot below, where movement along the X axis between subplots represents changes in GDP year to year, while movement along the Y axis represents changes in LEABY from year to year.

Germany and Chile, which have the highest LEABY values, have much lower GDP than countries like China and the United States. The United States LEABY trails Chile and Germany but surpasses Mexico, China, and Chile. Zimbabwe has extremely low GDP and also low LEABY values.

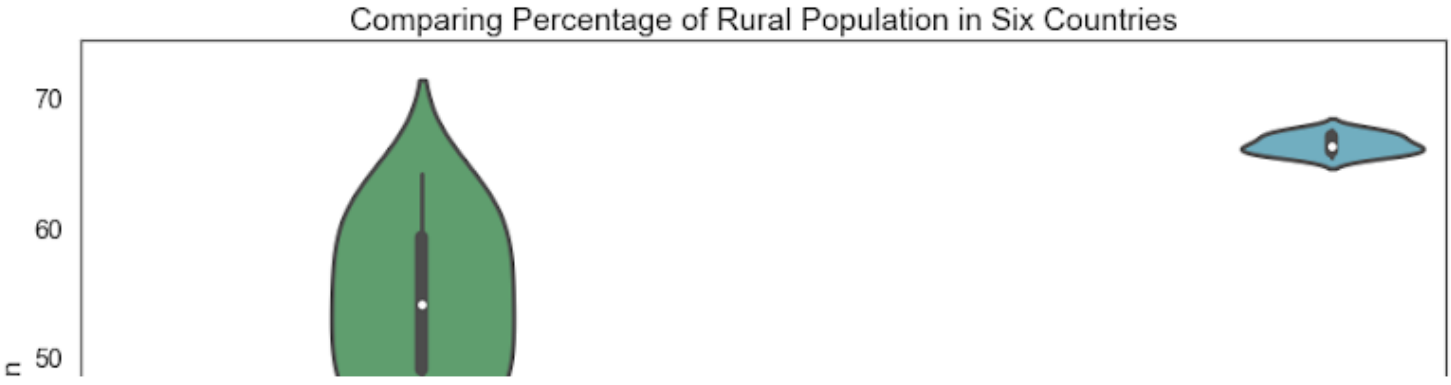
In terms of changes over time, GDP strongly increases in the United States of America and China over the 15 years under study, but despite rapid gains in GDP, these countries see only modest gains in LEABY. Germany and Mexico show more limited gains in GDP, but LEABY gains similar to those of the United States of America. Although Zimbabwe's GDP is very low and thus changes are hard to see in plots on this scale, LEABY increases markedly from 2005 to 2015 in this country.

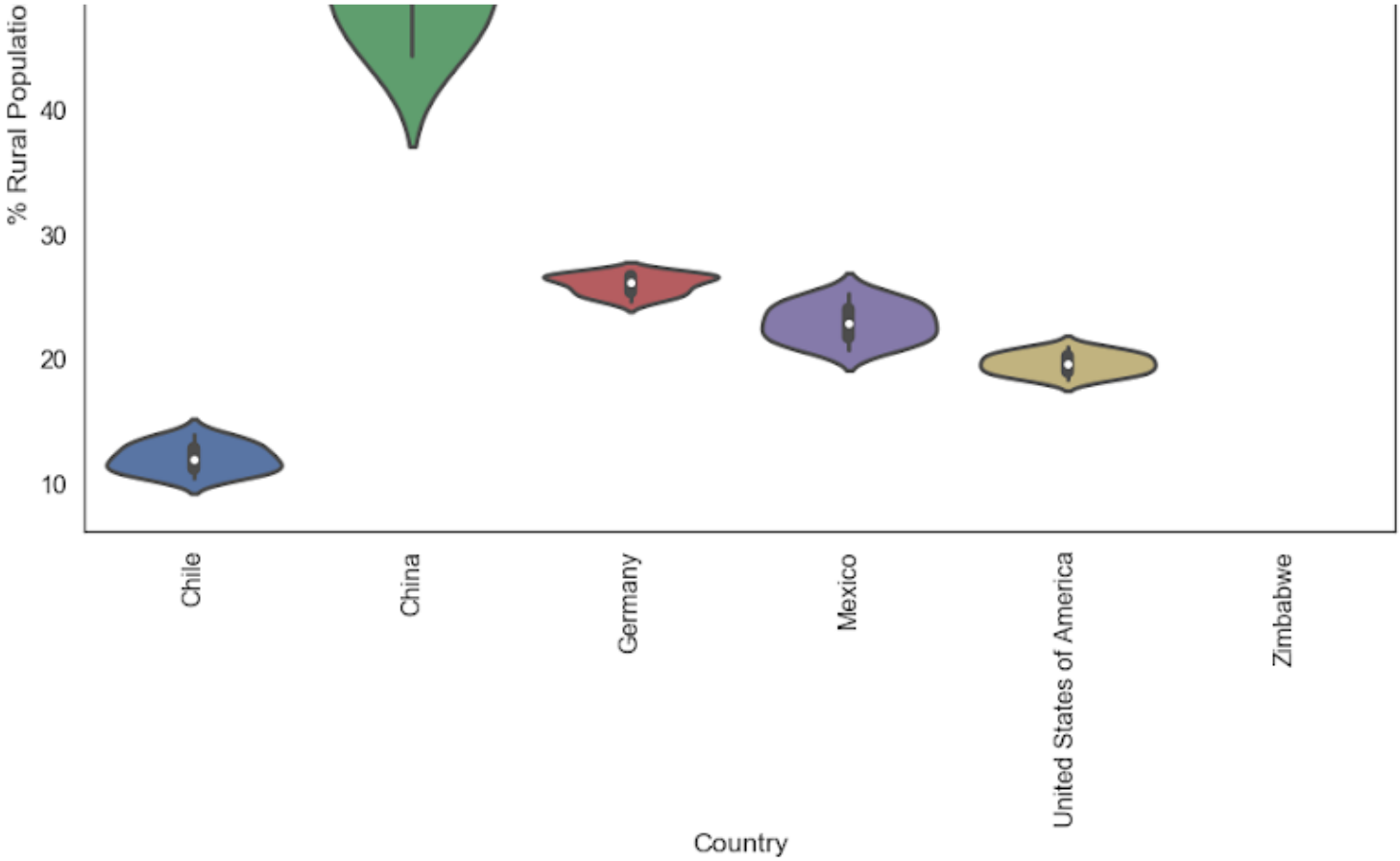
According to this data, LEABY and GDP do not seem to be correlated.



If not GDP, then what factors ARE correlated with Life Expectancy (LEABY) values?

It seems to make sense that GDP would be correlated with Life Expectancy, as richer countries would have more resources to address factors influencing life expectancy. However, this does not seem to be the case. However, GDP does not track how a country's wealth is distributed at the individual level, nor what the GDP is used for at a national level. Perhaps looking at a measure that tracks population living patterns and movement between rural and urban areas will show more correlation with LEABY values.





What % of the total population in each of these countries is living in rural areas?

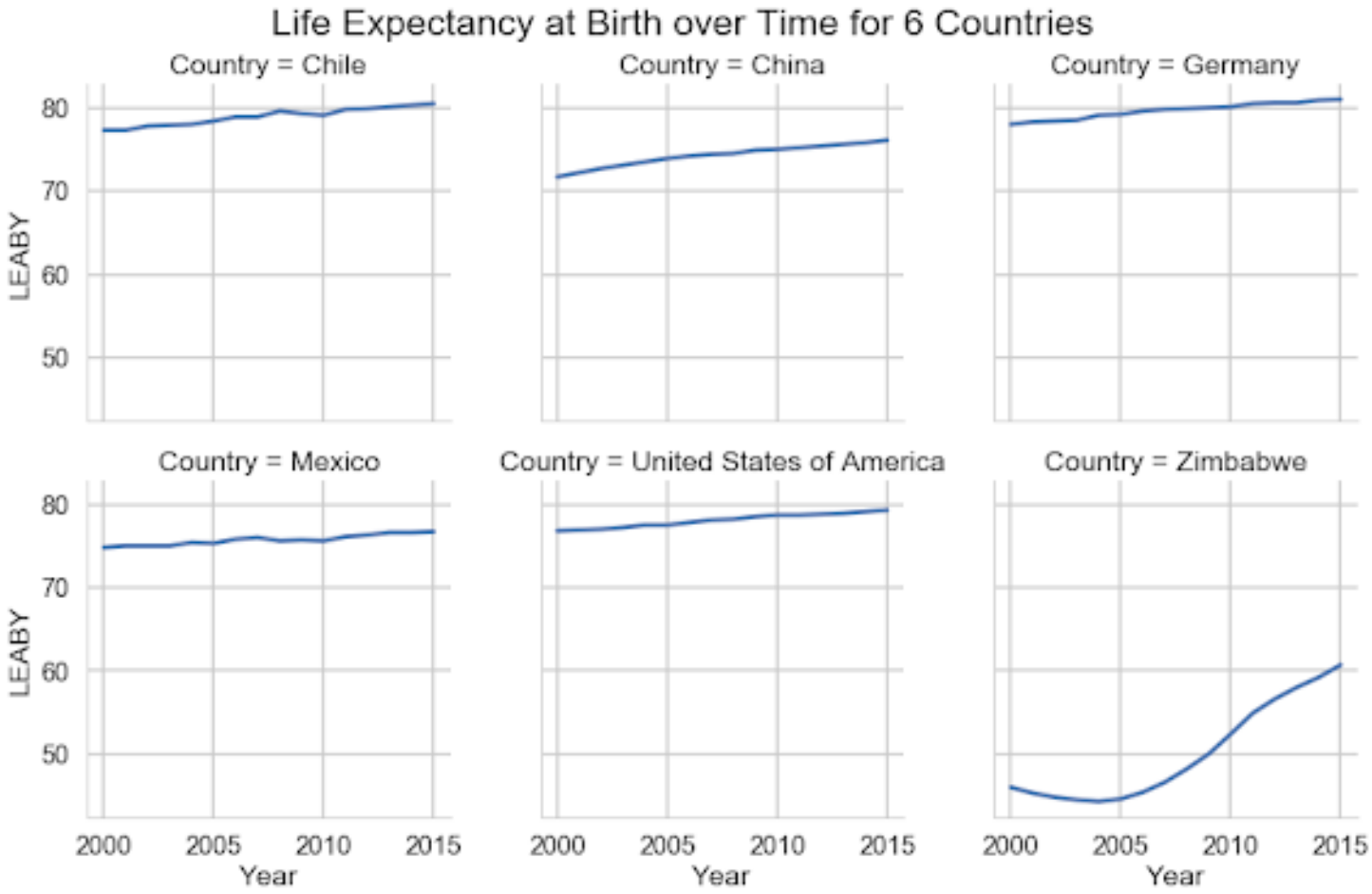
Another hypothesis is that expedient access to health care may reduce death rates and thus be correlated with Life Expectancy (LEABY) values. Though not a true measure of access to health care, perhaps the percentage of the population living in rural areas could be negatively correlated with LEABY, as rural populations may be further from hospitals and treatment available there. To look at this factor, I pulled data on the rural population as a percentage of the total population from the World Bank's Data Portal. These are estimates based on the United Nations Population Division's World Urbanization Prospects: 2014 Revision.

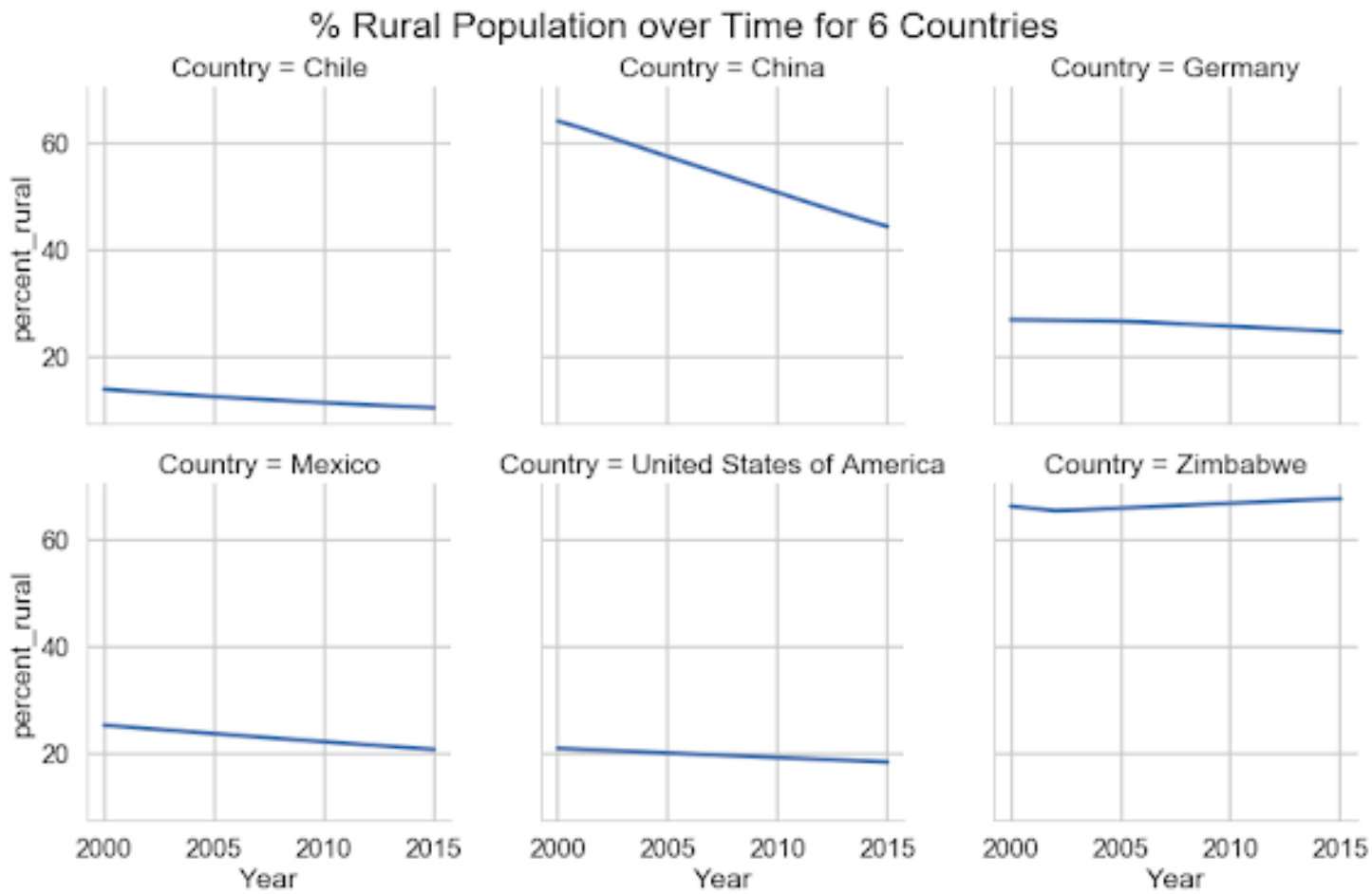
The data show that Zimbabwe has the largest rural population, and China the next largest population. Germany, Mexico and the United States are next in decreasing order, with Chile having an exceptionally urban population.

In the plots below it is easy to compare Life Expectancy at Birth with % Rural Population over time.

How does LEABY compare to the % Rural Population in each country, and how do those relationships change over time?

In the plots below, Chile has both a high LEABY and a low % Rural Population, as we might expect from our hypothesized relationship, and the trends seem to track opposite of each other. The correlation seems to hold for the China, Mexico and the United States of America. However, Germany seems to have a high LEABY and yet also a high percentage rural population. This may be a function of how "rural" is defined however, as a closer look at actual distances between rural communities and urban centers with hospitals may mean that in this case assuming rural communities have less access to hospitals may not be valid. Zimbabwe is the only country in which rural populations grew over the time period under study, but despite populations moving out of urban centers, the LEABY rose quite a bit over the same period. However, political unrest, widespread famine, and disease outbreaks during the period under study led to a humanitarian crisis in the area according to the World Health Organization. This could certainly explain some of the patterns in this country's LEABY, GDP, and % rural population statistics. In China, a large move towards urbanization was met with only modest gains in LEABY. It is possible that negative health effects of urban industrial environments, including high levels of air pollution could counteract any benefits of increased access to hospitals in some cases, especially under conditions of rapid industrialization.


Life Expectancy at Birth over Time for 6 Countries
Country = Chile
Country = China
Country = Germany
Country = Mexico
Country = United States of America
Country = Zimbabwe
LEABY
80
70
60
50
2000
2005
2010
2015
Year


% Rural Population over Time for 6 Countries
Country = Chile
Country = China
Country = Germany
Country = Mexico
Country = United States of America
Country = Zimbabwe
percent_rural
60
40
20
2000
2005
2010
2015
Year

Conclusion: Life Expectancy at Birth seems more correlated to % Rural Population than to GDP

However, the relationship between Life Expectancy and both GDP as well as % Rural population is complicated and LEABY is likely to be influenced substantially by historical factors like rapid industrialization (as in China) and humanitarian crises (as in Zimbabwe). In addition, we should consider the quality of data used in each case, and that 1) it is possible for some of these data sets to be manipulated at a national level for political reasons and 2) in low-GDP countries the completeness and quality of data is likely to be lower than in countries with more income to invest in vital record collection and maintenance.

Beyond the specifics of this capstone project, I've enjoyed this Codecademy course a lot and am feeling confident in my abilities to use Python to create useful data visualizations to support my work!