Emily Lutz (eel52)

Min Xia (mx76)

Jacques Diec (jpd276)

INFO/CS 3300

02 March 2015

Project #1 Report

**Description & Variables**

In this project, we were interested in measuring the effects of different variables on overall life expectancy in every country (and some sovereign territories) in the world. The overall life expectancy for each country and territory is our test variable. It is the mean of the overall life expectancy data for males and females from all age groups, demographics, and income levels. We introduced three control variables: Gross Domestic Product (GDP), ratio of national healthcare expenditure (public and private) to GDP, and total population.

**Data Collection**

We used 2012 data from the World Bank ([data.worldbank.org](http://data.worldbank.org/)). Data included overall life expectancy (test variable), ratio of healthcare expenditure (public and private) to GDP (control variable), population (control), and GDP (control). Each variable was reported in a different data set. In total, we used six data sets (.json), extracting only from columns that contained data from 2012.

One of the challenges in the data collection (and interpretation) process was handling null fields. These issues occurred because the World Bank did not report statistics for certain countries and territories in 2012. We resolved this by looking up the same data for those territories in other sources such as the Human Development Index (HDI). If no such data could be found, we averaged the data from 2011 and 2013. We then updated the null entries in our JSON files with the information we acquired.

The other challenge we faced was filtering data from sources that did not qualify as countries or territories, such as OPEC. We did not include them in our representation because the size of our data points was a function of population size. Since these organizations entail many countries, but themselves are not countries, their presence could lead to misinterpretations of the data. For instance, China, with a population of approximately 1.3 billion people, should be represented by the largest circle in the graph. However, the largest data point (before filtering) would be one containing the World population data (since China is a subset of the world). We were only interested in analyzing data from countries and territories.

The last (and greatest) challenge was processing data from multiple JSON files. We merged the datasets based on the common entry key that was the country name. We used data from six JSON files (healthExpenditure.json, population.json, GDP.json, lifeExpectancy.json, countryposition.json and world.json). The last two were used exclusively for the map.

In the first image, we constructed a world map, using the world.json file. We then acquired the central x and y coordinates of the countries and territories from countryposition.json to determine the position of our data points. Furthermore, we scaled the radii of the data points in correspondence to population and color in correspondence to life expectancy.

**Data Representation & Scaling**

In the representation of our data, we applied a linear scale to each variable except population, which was scaled by square root approximation. In constructing the axes for the graph, we scaled the first value of the range to height of our svg element for the (vertical) y-axis and the domain to the lowest value (rounded down) and the highest value (rounded up) in the life expectancy data set. The range for the (horizontal) x-axis had the values in reverse, with the domain entries being the lowest and highest values (rounded down and rounded up respectively) from the health care expenditure data set. The x-coordinate of the data points were scaled logarithmically, with the value being a function of expenditure. The y-coordinate of the data points were scaled linearly, with the values being a function of life expectancy.

Population and health care expenditure were the only variables we scaled non-linearly because we were interested in using population to determine the radius of each data point and a linear scale would result in extreme disparities in the size of the data points, and health care expenditures were measured as percentages of GDP.

**Interpretation & Analysis**

The first diagram (world map) depicts a geographic representation of life expectancy across regions. The highest life expectancies occur in Europe, North America, and Australia, where overall life expectancy is at least 74 years of age. Moreover, developing countries such as China, India, and Brazil have above-average to high life expectancies, with life expectancies of 60-73 years of age. The lowest life expectancies occurred in Africa and parts of Asia and Latin America, with 59 years of age or under. Average life expectancy in Africa is significantly lower than average life expectancy in other continents.

The visualization of our data suggests that our test variable has a moderate positive correlation with our main control variable. Life expectancy appears to increase at a decreasing rate as the ratio of healthcare expenditure to GDP increases. Furthermore, countries and territories with low populations and moderate-to-high GDP (e.g. Australia, Canada, Finland) tend to spend less on healthcare overall (despite popular perception), but are able to maintain high overall life expectancies. Other countries with high life expectancies like the United States tend to spend more on healthcare, although the overall proportions may be skewed due to fact that most of that spending is private. Developing countries like Sierra Leone (West Africa) tend to have low life-expectancy, despite high health care spending. However, the country has historically been plagued by war, disease, corruption, and poverty. Sierra Leone was also the source of the 2014 Ebola Outbreak. It is difficult to represent political factors on a graph because they are categorial variables that cannot be quantified objectively. Thus, it is best to provide written explanations (or annotations) for outliers.

Sources

GDP

<http://data.worldbank.org/indicator/NY.GDP.MKTP.CD?display=default>

Healthcare Expenditure to GDP (%)

<http://data.worldbank.org/indicator/SH.XPD.TOTL.ZS/countries/1W?display=default>

Life Expectancy from Birth

<http://data.worldbank.org/indicator/SP.DYN.LE00.IN?display=default>

Total Population

<http://data.worldbank.org/indicator/SP.POP.TOTL?display=default>

Country position

<https://opendata.socrata.com/dataset/Country-List-ISO-3166-Codes-Latitude-Longitude/mnkm-8ram>