Getting the Research Project Started

(Data Management and Visualization Week 1 Assignment)

STEP 1: Choose a data set that you would like to work with.

Gapminder is a non-profit venture that seeks to promote sustainable global development and achievement of the United Nations Millennium Development Goals. It seeks to increase the use and understanding of statistics about social, economic, and environmental development at local, national, and global levels.

The social and economic aspects of the data provided by Gapminder through their dataset, made me interested in it and I have decided to choose the same for further analysis and research.

Gapminder contains data for all 192 UN members, aggregating data for Serbia and Montenegro. Additionally, it includes data for 24 other areas, generating a total of 215 areas. GapMinder collects data from a handful of sources, including the Institute for Health Metrics and Evaulation, US Census Bureau's International Database, United Nations Statistics Division, and the World Bank.

The dataset provided contains the following parameters:

Variable Name	Description of Indicator	Main Source
incomeperperson	2010 Gross Domestic Product per capita in constant 2000 US$. The inflation but not the differences in the cost of living between countries has been taken into account.	World Bank Work Development Indicators
alcconsumption	2008 alcohol consumption per adult (age 15+), litres. Recorded and estimated average alcohol consumption, adult (15+) per capita consumption in litres pure alcohol	WHO
armedforcesrate	Armed forces personnel (% of total labor force)	Work Development Indicators
breastcancerper100th	2002 breast cancer new cases per 100,000 female. Number of new cases of breast cancer in 100,000 female residents during the certain year.	ARC (International Agency for Research on Cancer)
co2emissions	2006 cumulative CO2 emission (metric tons), Total amount of CO2 emission in metric tons since 1751.	CDIAC (Carbon Dioxide Information Analysis Center)

femaleemployrate	2007 female employees age 15+ (% of population). Percentage of female population, age above 15 that has been employed during the given year.	International Labour Organization
hivrate	2009 estimated HIV Prevalence % - (Ages 15-49). Estimated number of people living with HIV per 100 population of age group 15-49.	UNAIDS online database
internetuserate	2010 Internet users (per 100 people) Internet users are people with access to the worldwide network.	World Bank
lifeexpectancy	2011 life expectancy at birth (years). The average number of years a newborn child would live if current mortality patterns were to stay the same.	1. Human Mortality Database, 2. World Population Prospects: 3. Publications and files by history prof. James C Riley 4. Human Lifetable Database
oilperperson	2010 oil Consumption per capita (tonnes per year and person)	BP
polityscore	2009 Democracy score (Polity). Overall polity score from the Polity IV dataset, calculated by subtracting an autocracy score from a democracy score. The summary measure of a country's democratic and free nature. -10 is the lowest value, 10 the highest.	Polity IV Project
relectricperperson	2008 residential electricity consumption, per person (kWh). The amount of residential electricity consumption per person during the given year, counted in kilowatt-hours (kWh).	International Energy Agency
suicideper100th	2005 Suicide, age adjusted, per 100,000. Mortality due to self-inflicted injury, per 100 000 standard population, age adjusted	Combination of time series from WHO Violence and Injury Prevention (VIP) and data from WHO Global

		Burden of Disease 2002 and 2004.
urbanrate	2008 urban population (% of total). Urban population refers to people living in urban areas as defined by national statistical offices (calculated using World Bank population estimates and urban ratios from the United Nations World Urbanization Prospects)	World Bank

STEP 2. Identify a specific topic of interest

After looking through the codebook for the Gapminder dataset, I have decided that I am particularly interested in the life expectancy parameter.

STEP 3. Prepare a codebook of your own (i.e., print individual pages or copy screen and paste into a new document) from the larger codebook that includes the questions/items/variables that measure your selected topics.)

Variable Name	Description of Indicator	Main Source
lifeexpectancy	2011 life expectancy at birth (years). The average number of years a newborn child would live if current mortality patterns were to stay the same.	1. Human Mortality Database, 2. World Population Prospects: 3. Publications and files by history prof. James C Riley 4. Human Lifetable Database

STEP 4. Identify a second topic that you would like to explore in terms of its association with your original topic.

On subsequent analysis, I could think of checking the association of life expectancy against urbanization of the country/area under consideration. So, my research question is :

- Is average life expectancy associated with urbanization?

STEP 5. Add questions/items/variables documenting this second topic to your personal codebook.

Variable Name	Description of Indicator	Main Source
urbanrate	2008 urban population (% of total). Urban population refers to people living in urban areas as defined by national statistical offices (calculated using World Bank population estimates and urban ratios from the United Nations World Urbanization Prospects)	World Bank

STEP 6. Perform a literature review to see what research has been previously done on this topic. Use sites such as Google Scholar (http://scholar.google.com) to search for published academic work in the area(s) of interest. Try to find multiple sources, and take note of basic bibliographic information.

I have done a quick exploration on existing literature available via Google Scholar using the key terms as “life expectancy” and “urbanization”. There are only very limited number of literature available that are closely related to these two aspects put together. I have listed the most relevant ones below:

1. Kim, Jong In, and Gukbin Kim. "Country-level socioeconomic indicators associated with healthy life expectancy: income, urbanization, schooling, and internet users: 2000–2012." Social Indicators Research 129, no. 1 (2016): 391-402.

   This study confirms the associations between healthy life expectancy (HLE) and country-level socioeconomic factors. The analysis is based on HLE data of 178 countries obtained from the World Health Organization. Data on country-level socioeconomic indicators were obtained from the World Bank database and the United Nations. The associations between socioeconomic indicators and HLE are assessed using Pearson’s correlation coefficients and multiple regression models. The findings show significant positive correlations between HLE and the following country-level socioeconomic factors: national income level, urban population, mean years of schooling, and internet users. Based on the results, country-level socioeconomic indicators seem to have an important effect on healthy life expectancy.

2. Eckert, Sophie, and Stefan Kohler. "Urbanization and health in developing countries: a systematic review." World Health Popul 15, no. 1 (2014): 7-20.

   In this work, there are eleven studies of the association between urbanization and the selected health indicators in developing countries met our selection criteria. Urbanization was associated with a lower risk of undernutrition but a higher risk of overweight in children. A lower total fertility rate and lower odds of giving birth were found for urban areas. The

association between urbanization and life expectancy was positive but insignificant. Common risk factors for chronic diseases were more prevalent in urban areas. Urban-rural differences in mortality from communicable diseases depended on the disease studied. The study concludes that there are several health outcomes that are correlated with urbanization in developing countries. Urbanization may improve some health problems developing countries face and worsen others.

STEP 7. Based on your literature review, develop a hypothesis about what you believe the association might be between these topics. Be sure to integrate the specific variables you selected into the hypothesis.

Based on the literature review, I could come up with the following hypothesis for further analysis using the Gapminder dataset.

Average life expectancy of a country/area is improving with urbanization

Based on the dataset selected “lifeexpectancy” parameter should be directly correlated with “urbanization” parameter.