CS 418 Data Science
Final Project

Brian De Villa, Katherine Misyutina, Matthew Jankowski, Abhi Shah

Proposal

Questions:

- How does child mortality rate relate to development in countries worldwide?
- How does child mortality of measles compare to other diseases?
- What kinds of diseases are highly linked to underdeveloped/developing countries compared to developed ones?

Hypotheses and Predictions:

Countries with lower child mortality rates have a higher Gross Domestic Product.
Countries that have had a large decrease in child mortality rates over the past 10 years have had a large increase in Gross Domestic Product.
Countries with low GDP have higher mortality rates.

We will view our predictions through our maps, plots, and correlation matrix.

Data Collection:

For our analysis, we are collecting data from three sources,

We are collecting data from an official government website by the World Health Organization under the category of mortality and global health estimates.

We are analyzing the mortality for the following diseases / conditions:

- Acute Lower Respiratory Infections
- Prematurity
- Sepsis
- Measles
- Injuries

We are analyzing data for the following countries (around 4-6 per continent):

Asia:
India
Philippines
Bangladesh
China
Saudi Arabia

Africa:
Uganda
Kenya
Ethiopia
Morocco
South Africa
Nigeria
Burundi

Europe:
Poland
Russia
Germany
Ukraine
Serbia
Albania

North America:
United States
Canada
Mexico
Dominican Republic
Guatemala
Haiti

South America:
Ecuador
Colombia
Brazil
Chile

Oceania:
Australia
New Zealand
Solomon Islands
Fiji

We are also collecting GDP data from the "Our World in Data" website that shows the GDP from the 1990s to 2017.

Finally, we are web-scraping the population for each country from Wikipedia.

Anticipated Challenges:

- Preparing the data is important to start analysis.
- Choosing what areas to concentrate on. That includes:
    - What diseases?
    - What countries?
    - What years?
- There may not be enough data to compare the mortality of measles with other diseases.
- Creating a correlation matrix / plots / maps as visuals to display our results.

What importance your project will have on the related field:

Analyzing and researching from the datasets we have, we can find whether our hypotheses were valid. Either way, we can use the results to determine the need various countries face in child mortality. If we show that certain countries with high child mortality rates have low GDPs, we can provide them with aid. We can also create a broader overview of GDP vs child mortality for doctors and world leaders.

Progress Report

1. Data
    a. All but one of the datasets were csv files. The population dataset was webscraped from Wikipedia. The five diseases had all deaths for the children from certain age groups from the years of 2000-2017. The GDP dataset has all GDP per capita for each country since 1990-2017. The Wikipedia Population Dataset has the population for the entire country from 1955-2050.

    b. The data and scripts can be found at: https://github.com/ashah244/CS-418-Project

    c. Please look at this link for data samples. Each folder has the uncleaned and cleaned datasets with n < 50. https://github.com/ashah244/CS-418-Project/tree/master/datasets

2. Collection Process
    a. We started out by asking what data is needed to answer our questions. We found the csv files online by searching for the data we were interested in. For country population, we used BeautifulSoup and Pandas to scrape Wikipedia, as in Lab 2.

    b. We wrote scripts in Python that put the data in the way we wanted. To do that, we restructured columns, took out unnecessary data rows, took care of missing data, and saved the cleaned version in csv format.

    c. Data preparation is crucial in order for it to be used for analysis. We had to make sure that the data was properly formatted. Filtering columns by country was a bit challenging, but we wrote a function to solve that.

3. Challenges and Observations

    Some of the initial countries we had picked were missing data from the key years we were interested in. Instead of changing the years we were looking at, we opted instead to switch that country out with another that was similar in economic status and population. We want to make sure that we can represent poor, developing, and wealthy countries. We also wanted to make sure we had countries from all continents.

4. Next Steps

We haven't started analyzing and comparing the data from the different datasets. So our next step will be to put the data together. We will import the cleaned csvs into a python file and use libraries such as pandas and eda to produce plots. We may use geopandas to show results on a map. We might also use a correlation matrix to see correlation. We will use these in our analysis.

5. Group Member Duties

- Abhi: Cleaned Disease population data.
- Brian: Cleaned GDP dataset.
- Katherine: Researched child mortality datasets and countries.
- Matthew: Cleaned population data.