MSDS 6372 Project 1

Kaggle Dataset: “Life Expectancy (WHO): Statistical Analysis on factors influencing life expectancy”.

Kevin Boyd & Shikha Pandey

Introduction:

The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries. The datasets are made available to the public for the purpose of health data analysis.

We are doing a statistical analysis on factors influencing life expectancy. This study focuses on immunization factors, mortality factors, economic factors, social factors, and other health related factors. Our

Lot of studies were done in the past on factors affecting life expectancy considering demographic variables, income composition and mortality rates. However, effect of immunization and human development index was not considered, and those studies were done on dataset of one year for all the countries. This study includes data from a period of 2000 to 2015 for all the countries and important immunizations like Hepatitis B, Polio and Diphtheria are some of the factors included.

Since the observations this dataset are based on different countries, it will be easier for a country to determine the predicting factor which is contributing to lower value of life expectancy. This will help in suggesting a country which area should be given importance to efficiently improve the life expectancy of its population.

Data Description:

Information in this dataset related to life expectancy, health factors for 193 countries has been collected from the WHO data repository website and its corresponding economic data was collected from United Nation website for the period 2000-2015.

This data set lists data for countries from 2000-2015 for various immunizations - Hepatitis B, Polio and Diphtheria, diseases - Measles, HIV/AIDS, health factors - Alcohol, BMI, mortality factors – adult mortality, infant deaths, under-five deaths, economic factors – GDP, percentage expenditure, Total expenditure, Income composition of resources and social factors – population, schooling, status, thinness 1-19 years, thinness 5-9 years

Initial look at the data shows most missing values are for population, Hepatitis B, GDP. We have decided to move forward with keeping all the data to do our Exploratory Data Analysis.

Exploratory Data Analysis:

Objective 1:

Objective 2: