Life Expectancy (WHO) Dataset
STAT 4355.001 Team 11 Project Proposal
Abraca-Data Team Members: Prachi Patel, Evan Meade, Alejandro De La Cruz
Link: Life Expectancy (WHO) | Kaggle

Goal:
This dataset containing **22 variables** and **2938 entries** is about the life expectancy in different countries. The dataset has gathered information from **193 countries**, which has been collected from the World Health Organization (WHO) data repository website, who keep track of health status and many other factors for all countries. Our goal with this project is to investigate the key factors that help improve a countries' overall life expectancy. We want to find out what happens to the life expectancy in a country when certain statuses change. For instance, if the amount of schooling a person gets in a country increases or decreases, how will that affect the life expectancy in the country? By understanding why a certain country has a high or low life expectancy, we can figure out how to improve life for them using the information found.

The Dataset:
Though this data was originally collected by WHO and the United Nations (UN), we sourced it from Kaggle. This was done for convenience, as a user had already downloaded and combined the data into a convenient .csv form. As a result, we will have more time for analysis as we will be spending less time hunting data.

As stated in the goal, this dataset contains 2938 observations of 22 variables. The data contains large spatial and temporal baselines; it includes observations of **193 countries ranging from 2000-2015**. Each country is additionally classified as "developed" or "developing," providing opportunity for comparative analysis of these two groups. While not all (country, year) pairs have complete data, approximately half of them do, and many more have partially complete data. This allows us to assess the effects of time and place on life expectancy.

The main content of the dataset is a collection of health and economic indicators that measure societal factors which are potentially relevant to predicting life expectancy. The health factors include mortality rates (adult, children, infant, etc.), disease rates (HIV/AIDS, polio, measles, etc.), and national medical averages (life expectancy, BMI, alcohol consumption, etc.). Together, these figures provide a snapshot of the leading ailments in a country and the average person's health. Meanwhile, the economic factors track GDP, government health expenditures, and population, as well as average income and education levels. These factors give insights into the resources available to the government as well as the average citizen to face the medical ailments captured by the health indicators. So, taken altogether, this dataset provides a coarse view of the resources invested by countries to address public health issues. As such, by analyzing it, we can assess what resources/ailments have the highest impact on life expectancies.

Indicate response/predictor variables:
For this analysis there are ten features within the dataset that we would like to choose to predict the **Life Expectancy**, which is our response variable. These ten features represent the predictors with the highest magnitude correlations with life expectancy ($R^2 > 0.4$). As such, they are promising candidates for developing a predictive model, but are obviously subject to change as our analysis progresses. We have performed some minor "*cleaning*" since there is about **1%** to **5%** of missing data in certain columns, where we just automatically dropped them for the sake of time. The amount of observations have significantly dropped by nearly half from **2938** to **1649 observations**. Later on, we would do some more preprocessing by doing imputation once we have chosen features to work with for this analysis.
Response: **Life Expectancy**
Potential Features/Predictors: **Schooling**, **GDP**, **Alcohol**, **BMI, Percentage Expenditure, Income composition of resources, HIV/AIDS, Thinness 10-19 years (and 5-9 years), Adult Mortality**.

- There is a strong positive correlation between **Schooling** and **Life Expectancy** of **0.73**. This could be due to education being well established and widespread in certain developed/wealthier countries. This means countries have less corruption, more developed infrastructure, access to healthcare, etc.

- Moderate correlation between **GDP** and **Life Expectancy** of **0.44**, which are likely due to the same reasons as the first bullet point.
- Within the correlation plot there is a moderate positive correlation between **Alcohol** and **Life Expectancy** of **0.40**. We could assume this is due to the fact that certain wealthier countries are able to afford the consumption of alcohol is more common among the wealthy.
- Moderate positive correlation between **BMI** and **Life Expectancy** of **0.54**. This could be due to having healthier lifestyles in developed/wealthier countries compared to developing countries.
- There is a moderate positive correlation between **Percentage Expenditure** and **Life Expectancy** of **0.41**. This could possibly be due to countries having the monetary funds where they could spend on healthcare which reflects on government spending on healthcare, resulting in the correlation of **Life Expectancy**.
- There is a strong positive correlation between **Income composition of resources** and **Life Expectancy** of **0.72**. A human development index of how developed a certain country is in terms of income composition of resources.
- Strong negative correlation between **Adult Mortality** and **Life Expectancy** of **-0.70**. Which indicates that there is a higher mortality rate across from both sexes which would result in having a lower life expectancy.
- Since there is no correlation for this feature, **Status**, it **could serve as an indicator** of which countries are developed or not, which could possibly give an accurate precision of the Linear Regression model.

Analysis Plan/Responsibilities:

To guide our analysis, we have developed the following questions we wish to explore. Some are inspired by the suggested questions provided by the dataset (linked above). Others are entirely developed by us. We may add or remove questions depending on how it pans out when we actually start working with the data.

1. Which predictors are most correlated with life expectancy? **ALL**
2. Is there a linear relationship between any of the predictors and life expectancy? Or is there some nonlinear relationship at play? **ALL**
3. Given these predictors, how accurately can one predict life expectancy with a linear model? **Alejandro**
4. Have the impacts of predictors changed over time or remained constant? (eg. was adult mortality always highly correlated with life expectancy, or is that a recent development?) **Evan**
5. Which countries over- or under-perform in life expectancy relative to what our linear model would predict? What might account for this difference? **Evan**
6. Are any of the predictors correlated with each other? If so, can the dimensionality of the dataset be reduced while retaining predictive power? **Alejandro**
7. To improve the lifespan of a country with low life expectancy(<65), should they improve on their healthcare expenditure? (*inspired by suggested question #2*) **Prachi**
8. In a developing country, what should they do to increase life expectancy versus a developed country? **Prachi**