**Introduction**

This case study will focus on the analysis of Life Expectancy Data, and on creating predictive models for life expectancy for any nationality.

For our first objective, we will build a model in order to identify key relationships along with all necessary testing and interpretations.

For our second objective, we will compare multiple models in order to develop a model that can highly predict life expectancy for any nationality.

**Data Description**

The data were provided by Dr. Turner for his MSDS 6372 class. The data can be found under Files, Project 1, Life Expectancy Data (1).csv within Digital Campus for course DS 6372 Applied Statistics. The data set is made up of 2938 observations and 22 explanatory variables.

\*No independence bc same country over time, repeated measures \* doesn’t take into account since the repeated measures of the same country are correlated , also year to year sampling will contribute to correlation and is outside the scope as it is related to Repeated Measures.

**Objective 1**

**Restatement of Problem**

As part of our first objective, we will build a model in order to identify key relationships along with all necessary testing and interpretations

**Preparation**

We start by reading in the data set, getting a sense of the overall data trend as we are interesting in in Life Expectancy as our response variable. We proceed to check all observations for missing points such as NAs. Upon visualizing the data and its missing data, we select variables that will be imputed based on a threshold of about 10% for missing data per variable. All variables, except for 5—Country, Year, Status, Life Expectancy and Hepatitis B—were selected to be imputed as they variables have numeric outputs and fall under the 10% threshold. We proceed to then remove 2 columns—Hepatitis B and Population from the data set and then remove any remainder NAs which were found our response variable—Life Expectancy. Research suggests that it is inadvisable to impute response variables, therefore we opted to only remove the <0.01% remaining NAs. Given this information we felt that the 19% for Hepatitis B and the 22% for Population were too great that we would risk biasing our overall data. Population also is related to the GDP within the data set. The GDP provided appears to be GDP per capita, as research would suggest, and GDP per capita is calculated by using GDP/Population, which population is a variable in our data set. GDP per capita and Population would be too closely related and would possibly attribute to covariance.

Talk about other variables that we will combine or address (ask Alex for this)