INTRODUCTION

Data Understanding: First, we created a data dictionary for the life expectancy data to understand what the indicator variables were measuring and the appropriate ways to investigate them.

Diagnosing and Cleaning: When diagnosis the data, we noticed that many of the variables had obvious error that appear to be related to an extraction error. Value that ended in 0 were truncated, drastically underrepresenting the true value. To address these errors, in addition to true missing values, we imputed any missing values and errors with averages based on similar time periods within the specific country in questions. We also decided to reduce the years variable from 2000-2015 to 2010-2015 because there were many error from the earlier years and the years between 2010-2015 would be more relevant to the current situation.

Exploring Relationships: After cleaning the data, we created a correlation matrix to find the most influential variables for life expectancy to use in the study and analyze further. From the correlation matrix we decided that Income composition of resources, Schooling, HIV/AIDS, and BMI were the variables that appeared to be the most deterministic of life expectancy, in addition to some moderately correlated features.

Analysis and Modeling: After determine the best indicators, we created a categorical representation of Life Expectancy to better understand trends in the quantitative factors, by comparing median measure across the categories. Furthermore, we aimed to build a parsimonious model to try to predict a countries life expectancy based on the identified factors.

RESULTS