Aaron Chumsky **Project Proposal 2**

**Title of the Project:** What is your life expectancy?

**Big Idea:** We will investigate a data set in order to determine the factors that best determine one’s life expectancy with the least amount of error. If we can determine what affects life expectancy most, perhaps people can begin to make lifestyle changes that can raise their expectancy or governments can invest more money in features that increase life expectancy.

**Data Description:** This data can be downloaded as a csv file from <https://www.kaggle.com/kumarajarshi/life-expectancy-who> with a free signup to the Kaggle website.

The dataset contains 2,938 observations and 22 variables.

* **Response:** Life Expectancy (labeled Life.expectancy in R) which is a variable indicating one’s life expectancy. The response is a quantitative variable with continuous values for age rounded to the nearest tenth.
* **Features of Interest:** All of the predictors have the potential of being significant determinants for life expectancy; the real question is to see which one’s are the most significant. Based on intuition, I would assume predictors such as development status, BMI, and mortality rates would be significant. It’ll be interesting to determine the significance for other variables such as years of education and government expenditure which will be seen once further testing has been done.
* **Missing Values:** Out of 2,938 observations, our response variable, Life Expectancy, only suffers from 10 missing values. The three variables that suffer from the most missing values are Hepatitis B with 553 missing values, GDP with 448 missing values, and Population with 652 missing values.
* **Readability into R:** The csv file downloads from the Kaggle website and reads into R very nicely. Observation numbers and column titles are included. The title of the columns are very clear in terms of what they represent, with descriptions of each variable on the Kaggle website.
* **Tidiness of Data:** Data is very tidy. There are two categorical variables, country and development status. Country takes on many different values and if it’s an important variable, it can perhaps be recoded into a dummy variable with many levels. Development status only has two values and can easily be recoded into a dummy variable with two values (1 for “developed” and 0 for “developing”). The other variables are quantitative either as numeric values or as percentages, so they are easily interpretable.

**Questions of Interest:** The primary question of interest is to determine which features are most significantly correlated with life expectancy (i.e BMI, prevalence of HIV, etc.). This problem is certainly more of an inference problem since we’re determining the effect of our independent variables on our response variable, as opposed to predicting the response variable (life expectancy). It wouldn’t make much sense to predict one’s life expectancy because life expectancy in itself is a prediction for how long one would live; we would need data on how long the individuals are actually living (which we don’t have) in order to determine predictive accuracy. This problem is also supervised given that this regression problem has a response variable that we’re working and using to make inferences. This is a regression problem because our response variable is continuous (life expectancy can take on any numerical value). It also may be more realistic to focus on variables that individuals or societies have control over i.e. BMI and government expenditure on health as opposed to Country.

**Interested Parties:** This can be seen on both an individual and a societal level. This can be of interest to individuals who want to increase their life expectancy and can do so by living a healthier lifestyle. For example, if BMI and alcohol consumption are significant predictors of life expectancy, they can start to watch their weight and drink less alcohol. On the other hand, this can also be of interest to governments who want to increase the life expectancies of their people. If expenditure on health is a very important predictor, they can start investing a lot more funds into health. If alcohol consumption is significant, they can perhaps start setting limits on how much people drink. If vaccination is important, they can increase its prevalence at health clinics. If life expectancy is looked upon at the government level, its impact can be very widespread.