**Slide Notes:**

**Slide 1: Intro Slide**

**Slide 2: Project Intro**

* Why is this project important
  + Obviously LE is important
    - We all want to
      * Live longer
      * Have our families around for longer
  + A lot of studies that look at life expectancy
    - Take into account demographic variables, income composition, mortality rates,
    - Or focus on individual LE, focusing on eating behaviors and healthy life style choices
  + Few have looked into features in this dataset like immunization, human development, gov investment in health services
* The problem/questions to be answered
* Regression Problem: Life Expectancy is a continuous target variable that ranges from about 36 to 89
* Data Set
  + Pulled from Kaggle
  + Health Factors –World Health Organization data repository
  + Economic Factors – United Nations Website

**Slide 3: EDA – Life Expectancy Histogram**

* Good starting point for understanding target variable
* A lot of information
  + Range: 36 – 89
  + The most common age for Life Expectancy is around 74 years old

**Slide 4: EDA – Development Status Boxplot**

* Good information about status feature (1 of 2 categorical features in dataset)
  + Average LE of developed countries is 10 years longer the developing countries
  + Even more telling: the minimum LE of developed countries is higher then the mean LE of developing countries.

**Slide 5: EDA – Immunization Heatmaps**

* Heatmaps of immunization percentages amount 1 year olds for Polio, and Diphtheria
* Found it interesting that immunizations for these two different diseases had very similar effects on life expectancy
  + Majority countries fell within the 95-100% immunization of 1 yos
  + For the countries/years with lower % of immunization rates in 1 yos, you do see a decline in LE
    - More true for diphtheria (less gradual slope)

**Slide 6: Splitting Data**

* Why Group – bc countries have multiple entries, most being from 2000-2015
  + Health advances that came about in 2000 inc LE, so I am looking at data post those health advances for more current LE data
* Group shuffle split to separate our test set from the rest of the data
* GroupKFold with 4 folds to create the training and validation sets
* Set sizes

**Slide 7: Preprocessing**

* Preprocessors:
  + OHE for categorical features – country and status
  + Standard Scalar for continuous features
    - SS vs MinMax because some of the continuous features (like GDP and population) were not reasonably bounded.
  + Data Shape
    - Column number grew significantly due to number of countries

**Slide 8: Missing Values**

* No missing values in either of the two categorical features, so I did not need to utilize Simple Imputer
* 44% of the entries have missing values
* 64% of features have missing values