Project 1 Presentation Script (Caleb)

INTRODUCTION SLIDE:  
  
Hello, everyone.

Welcome to Life Expectancy. Our group presenting today includes Stephanie Duarte, Steven Cox, and myself (Caleb Thornsbury). We are excited to share our data analysis of the Life Expectancy dataset with you today.

In this presentation, we'll be diving into EDA/Data Analysis of our dataset. We'll explore key relations between life expectancy and its predictors and along the way we’ll provide insights, strategies, and actionable takeaways to improving model performance while limiting confounding variables and choosing the most representative predictors.

TABLE OF CONTENTS SLIDE:

In this section, we'll provide an overview of the presentation and introduce the question we'll be addressing. We will also state the objectives for this presentation and go into detail over the process of EDA/Data analysis, which includes cleaning, transformation and interpreting the data. The last section of this presentation will be going over the big picture/using the EDA to explain the patterns we found, and models we created for this dataset.

DEFINING WHY SLIDE:

One of the first reasons to do this analysis is Problem Identification. This in tales examining the completeness, accuracy of the data, and disparities in data collection methods across regions or time periods. The “problem” we are trying to solve in this data is, are there any correlations between all variables/predictors and life expectancy. Just to name a few, this dataset includes country, schooling, and HIV/AIDS as variables. Challenges may arise from the need to account for confounding variables. Some of these confounding variables not found in this dataset include healthcare access, environmental conditions, and lifestyle behaviors.

Pattern recognition involves identifying trends, correlations, and anomalies that can provide insights into factors influencing life expectancy. This often begins with exploratory data analysis. Uncovering patterns visually through techniques such as scatter plots, box plots, and heatmaps are good places to start. More advanced statistical methods include regression analysis, clustering, and machine learning algorithms.

Continuous improvement involves the process of EDA. This can be described as refining data collection, enhancing analytical techniques, and updating models to better understand and predict factors influencing life expectancy. The EDA process begins with regularly evaluating the quality and completeness of the data, addressing any inconsistencies or missing information to ensure accuracy and reliability. After the first basic model we then employ advanced statistical and machine learning methods to uncover new insights and refine existing models for predicting life expectancy. This may include incorporating additional variables, optimizing model parameters, and validating results using bootstrapping methods. By continually refining the dataset and analysis methods, we can identify emerging trends, address evolving challenges, and develop better models for predicting and improving life expectancy.

MAKING OBJECTIVES SLIDE:

The first objective we have for this presentation is to build a regression model that will help us Identify the key relationships between the response variable, Life expectancy, and the other variables/predictors in the dataset. We will then interpret the relationship between these predictors and the response variable.

The second objective of this presentation is to build onto the first model. We will take the predictors that have the most influence and create 3 additional models to compare. The first model we will make is a Feature selection model using glmnet, the second model we will use is KNN, and lastly, we will use bootstrapping.

BOOTSTRAPPING SLIDE:

Bootstrapping is a resampling technique used in statistics to estimate the sampling distribution of a statistic by randomly selecting observations from the original dataset with replacement. This allows us to create new datasets of the same size as the original, mimicking the process of random sampling. Bootstrapping enables us to calculate confidence intervals and conduct hypothesis tests without relying on theoretical assumptions about the underlying population distribution. Instead, we use the distribution of the bootstrap statistic to make inferences about population parameters.

One of the upsides of Bootstrapping is that it is non-parametric, meaning it does not assume any specific distributional form for the data. This makes it particularly useful when dealing with complex or non-standard data distributions.

K-Nearest Neighbors regression is an intuitive algorithm for predicting continuous values. By leveraging the concept of similarity between data points, KNN regression provides a flexible and interpretable approach to machine learning. However, its performance depends heavily on the choice of parameter K and the nature of the data.

GLMNET regression is a powerful and versatile technique for regression modeling, offering the advantages of both Lasso and Ridge regularization. By automatically selecting important features and balancing the trade-off between bias and variance, GLMNET provides an effective tool for building predictive models in a wide range of applications.