**Afternoon Lab: Data Handling, Visualization, and Linear Modeling (4 hours)**

**Target Audience:** This lab is designed for students with no prior experience in statistical programming or data analysis.

**Learning Objectives:**

* By the end of this lab, you will be able to:
  + Describe the benefits of using statistical programming environments for public health data analysis.
  + Load and explore public health datasets using R.
  + Create various data visualizations to understand data distributions, relationships, and outliers.
  + Build a simple linear regression model in R to predict a public health outcome.
  + Diagnose potential issues with a linear regression model.
  + Apply model selection techniques to improve the performance of a linear regression model.
  + Gain a basic understanding of regularization techniques in linear regression.

**Software:** We will be using R, a free and open-source statistical programming language, for this lab.

**Materials:**

* A computer with R and RStudio (integrated development environment) installed.
* Downloadable public health data files (links will be provided).
* Lab manual with step-by-step instructions and code examples (provided below).

**Lab Schedule (approximate times):**

* **Introduction (15 min):**
  + Introduction to statistical programming and R for public health data analysis.
  + Overview of the lab structure and learning objectives.
* **Setting Up the Environment and Data Exploration (60 min):**
  + Hands-on installation of R and RStudio (if not already installed).
  + Introduction to the R interface and basic commands.
  + Loading and exploring a public health dataset (e.g., dietary habits and obesity rates) using R functions for data manipulation and summary statistics.
  + Learning resources for finding and accessing public health data.
* **Data Visualization (60 min):**
  + Introduction to data visualization concepts and their importance in data analysis.
  + Creating scatter plots for exploring relationships between variables.
  + Generating histograms to understand data distributions.
  + Understanding box plots and identifying potential outliers.
  + Practical exercises in creating different data visualizations in R using relevant packages like ggplot2.
* **Implementing Linear Regression (60 min):**
  + Introduction to linear regression concepts and their application in public health research.
  + Building a simple linear regression model in R to predict a public health outcome variable (e.g., obesity rates) based on another variable (e.g., dietary habits).
  + Examining the model output and interpreting coefficients.
  + Introduction to residual plots and their use in diagnosing model assumptions.
  + Hands-on exercise in building and interpreting a linear regression model in R.
* **Model Selection Techniques (30 min):**
  + Understanding the importance of model selection criteria for improving model performance.
  + Overview of common model selection techniques like AIC and BIC.
  + Practical demonstration of applying these techniques in R to select the best model for a given dataset.
* **Regularization Techniques (30 min):**
  + Introduction to the concept of regularization for reducing model complexity and overfitting.
  + Overview of different regularization techniques like L1 and L2 regularization.
  + Hands-on implementation of regularization techniques in R for improving a linear regression model.
  + (Optional) Brief discussion on the benefits and trade-offs associated with different regularization approaches.
* **Wrap-up (15 min):**
  + Q&A session for any questions or clarifications.
  + Wrap-up summarizing key takeaways from the lab.
  + Introduction to follow-up resources for further learning.