**🔹 Research Question:**  
Can we predict the prevalence of cardiovascular disease in California counties based on environmental pollution and socioeconomic factors such as air quality, drinking water quality, housing burden, poverty, and linguistic isolation?

**🔹 Data Sources:**  
The primary dataset for this analysis is the **CalEnviroScreen 4.0** dataset provided by the California Office of Environmental Health Hazard Assessment (OEHHA), which includes over 80 variables for 8,000+ census tracts in California. The dataset contains environmental indicators (e.g., ozone, PM2.5, diesel particulate matter), population health data (e.g., asthma, low birth weight, cardiovascular disease), and demographic factors (e.g., poverty rate, educational attainment, housing burden).  
📄 Source: [CalEnviroScreen 4.0 Data](https://oehha.ca.gov/calenviroscreen/report/calenviroscreen-40)

**🔹 Techniques to Be Used:**

1. **Exploratory Data Analysis (EDA)**
   * Correlation heatmaps, geospatial choropleths, and distribution plots
2. **Multiple Linear Regression**
   * Predicting cardiovascular disease rates from multivariate predictors
3. **LASSO Regularization and Sequential Feature Selection**
   * For model simplification and identification of key features
4. **K-Fold Cross-Validation and GridSearchCV**
   * To evaluate and optimize model performance
5. **Error Metrics**
   * Model accuracy will be evaluated using **Mean Absolute Error (MAE)** and **Huber Loss** to ensure robustness to outliers
6. *(Optional)* **PCA**
   * For dimensionality reduction if necessary due to multicollinearity or visual analysis
7. *(Optional)* **K-Means++ Clustering**
   * To identify similar census tract profiles based on environmental and social burden

**🔹 Expected Results:**

* A well-performing regression model that predicts cardiovascular disease prevalence based on both environmental burden and demographic vulnerability
* Identification of the most significant predictors (e.g., PM2.5 levels, drinking water contamination, housing burden)
* Insightful visualizations including maps of high-risk zones and feature importance plots
* Clear explanation of how much variance in cardiovascular health outcomes can be attributed to environmental and social factors

**🔹 Why This Question Is Important:**  
Cardiovascular disease remains a leading cause of death in the U.S., and environmental injustice plays a key role in increasing health risks for vulnerable populations. Communities exposed to high pollution levels and suffering from poverty, housing burden, or linguistic isolation often face compounded risks. By analyzing the intersection of environment and health, this project will provide **actionable insights** to public health officials and policymakers. These findings can help allocate resources more equitably, advocate for environmental reforms, and develop early intervention strategies in high-risk areas—ultimately improving community health outcomes across California.