## Data and Code

All analyses were conducted in Python (v3.11) using pandas, numpy, matplotlib, seaborn, statsmodels, and scikit-learn. Data cleaning, integration, and regression modeling steps were implemented in Jupyter Notebook (cic\_challenge\_regression.ipynb).  
  
**Workflow**

1. **Data Integration:** Public data from CDC (vaccinations, PLACES distress), the KFF Covid-19 Dashboard “Build a Custom State Report” tool (which draws from KFF and other credible sources), BLS (unemployment), BEA (gross state product), and Census were merged by state and year (2020–2023).
2. **Feature Engineering:** Poverty-adjusted measures, distress prevalence, and annual vaccination coverage were derived.
3. **Modeling:** Before regression, I calculated variance inflation factors (VIF) to assess multicollinearity, given the overlap among socioeconomic indicators. I then ran ordinary least squares (OLS) to establish baseline relationships and statistical significance, and least absolute shrinkage and selection operator (LASSO) regression to penalize overfitting, reduce redundancy, and highlight the most stable predictors.
4. **Validation:** I compared models on explanatory fit (R²), coefficient stability under regularization, and the degree to which predictors retained significance once shrinkage was applied.