Predicting Health Outcomes from Environmental Pollutants

Health Outcomes Predicted

Asthma

Cancer

COPD

Congestive Heart Failure

Kidney Disease

Stroke

Environmental Data Used

Particulate matter 2.5 level in air

Ozone level in air

Diesel particulate matter level in air

Air toxics cancer risk

Air toxics respiratory hazard index

Traffic proximity and volume

Percent 1960 housing (lead paint indicator)

Proximity to National Priorities List (NPL) [superfund] sites

Proximity to Risk Management Plan (RMP) facilities

Proximity to Treatment Storage and Disposal facilities

Indicator for major direct dischargers to water

Exploratory Data Analysis

COPD Heatmap with Environmental Factors
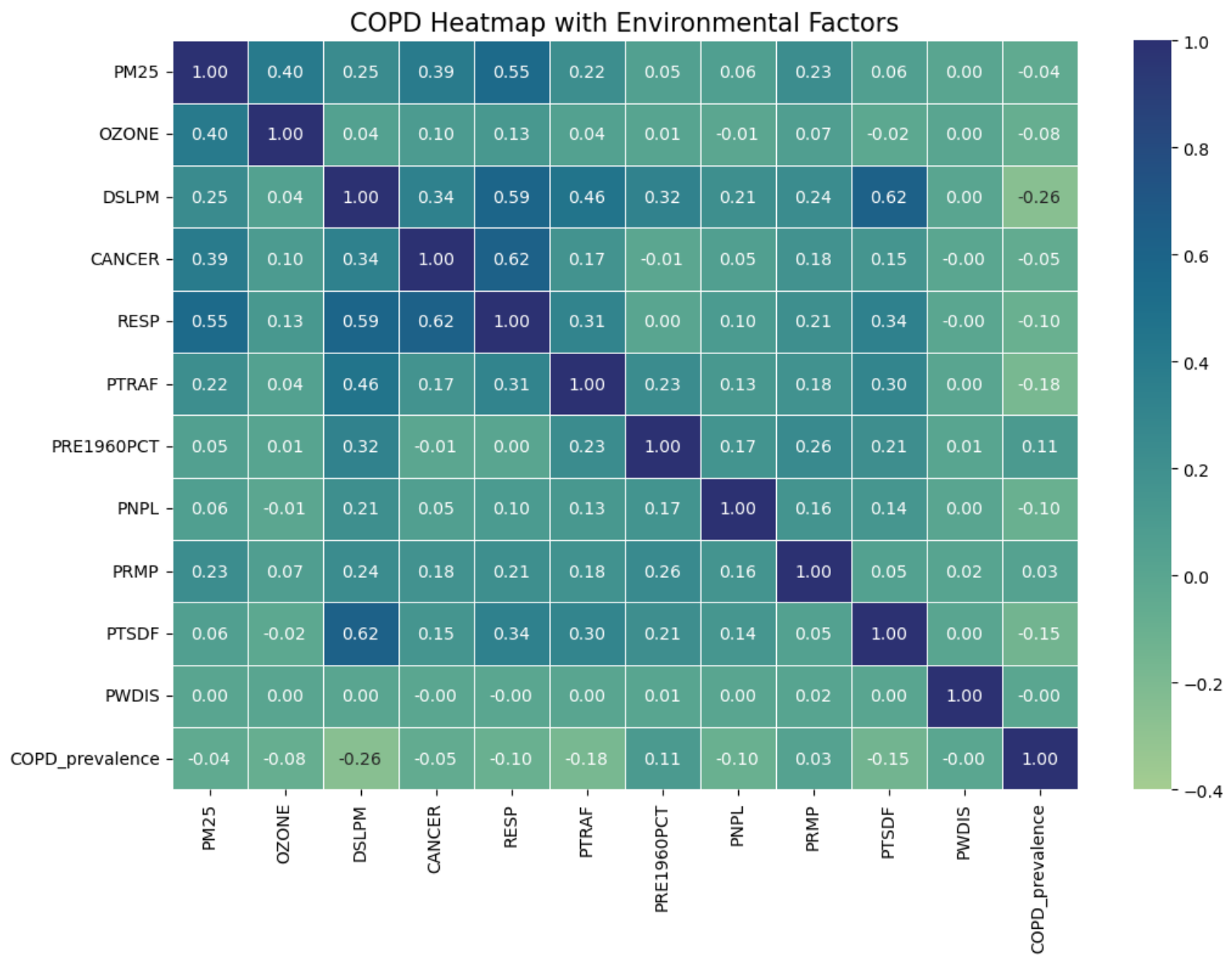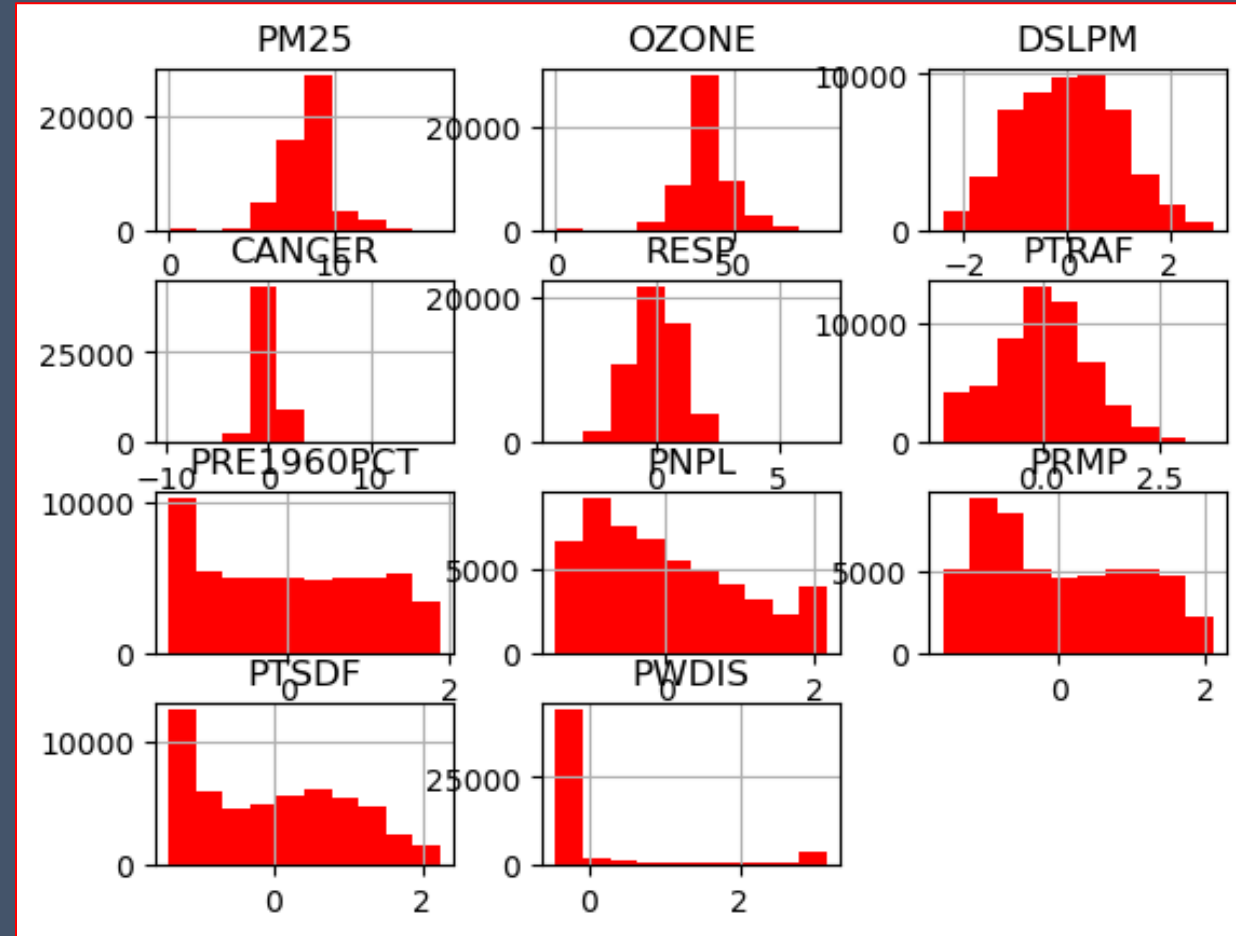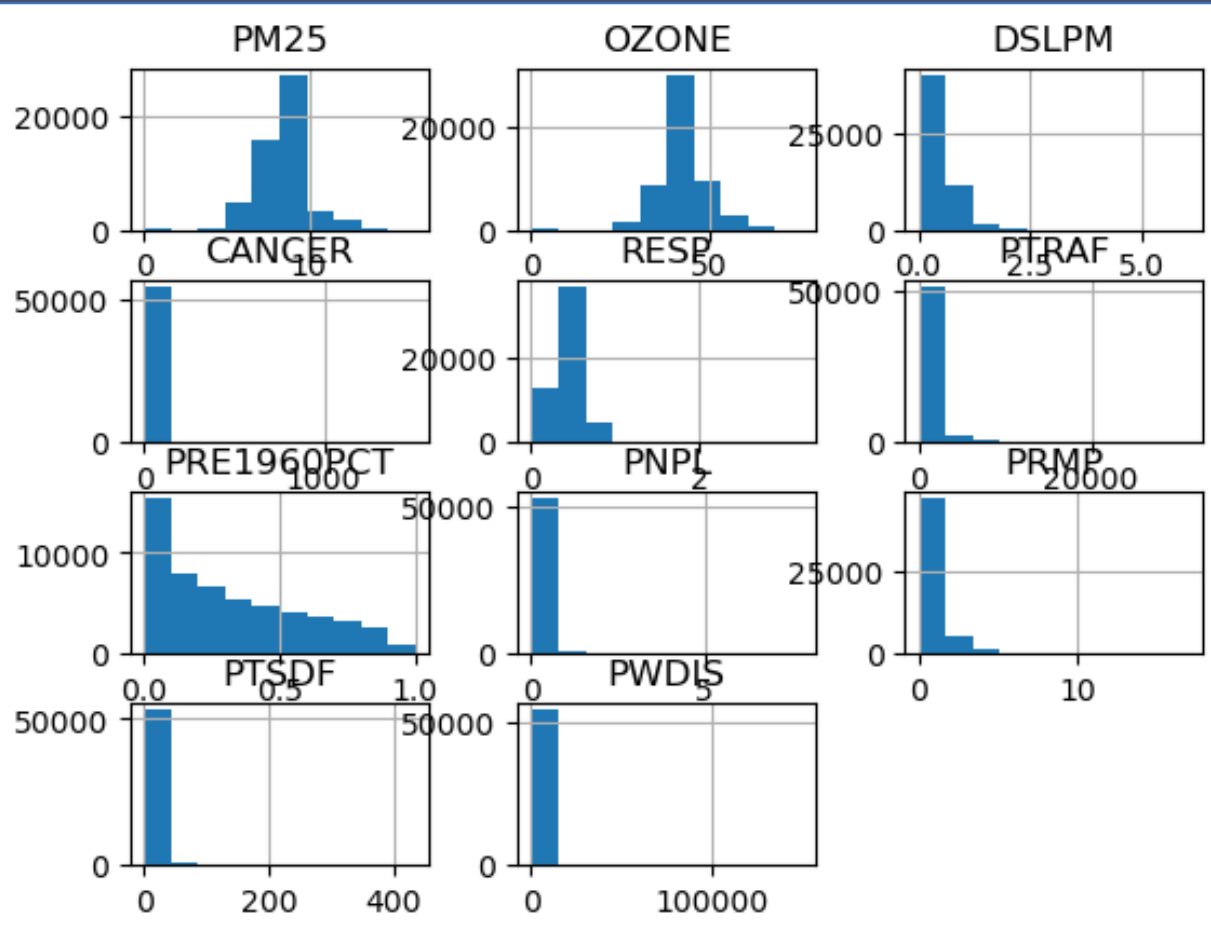
Normalizing Data

Types of Regression Modeling

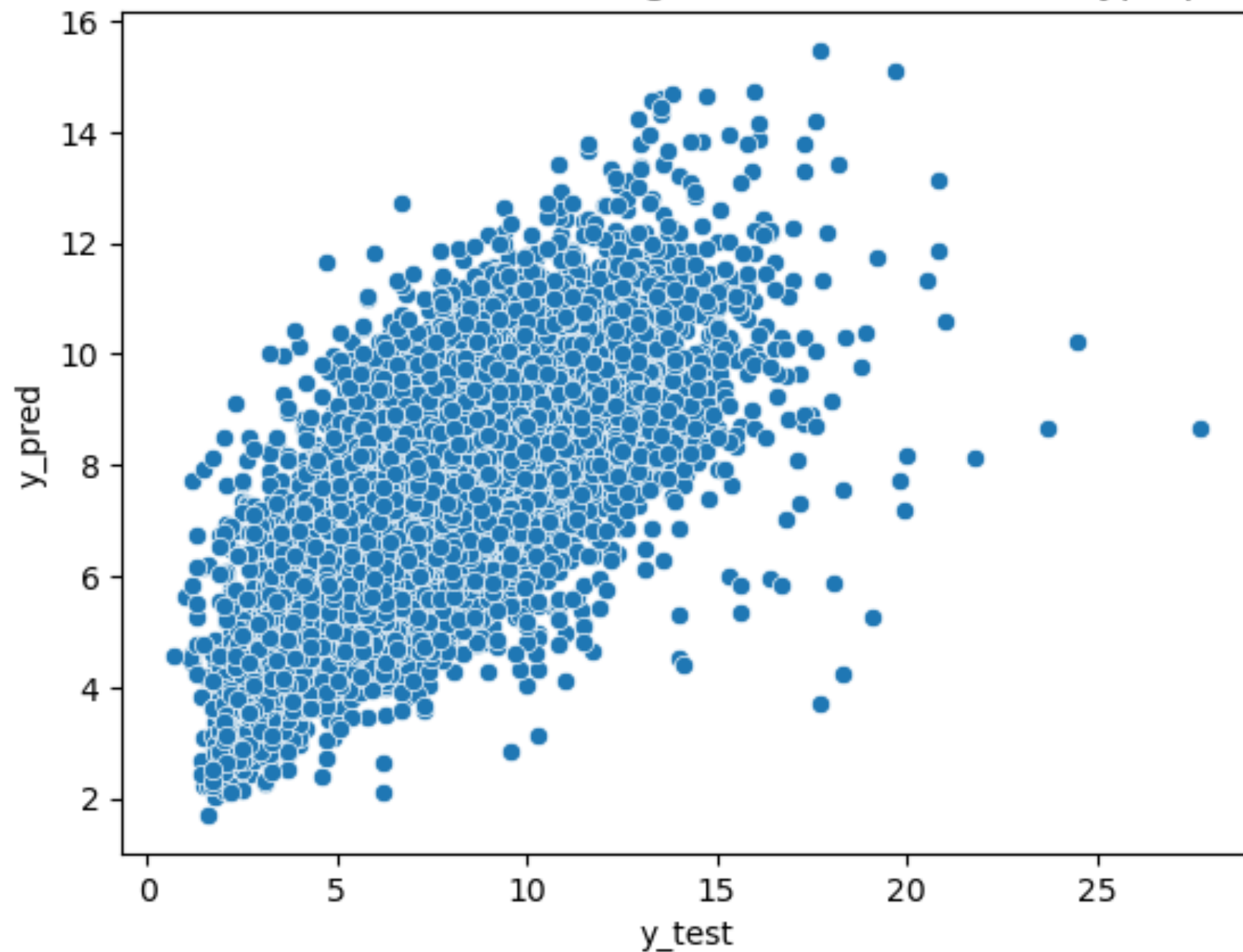Multiple OLS
Random Forest
AdaBoost
Gradient Boost

XGBoost

LightGBM

SVM

Predicted vs. Actual Values for COPD Using Random Forest with Hyperparameter Tuning

	Best Model	Type of Tuning	R^2
COPD	Random Forest	Bayesian	0.568
Asthma	XGBoost	Bayesian	0.554
CHD	Random Forest	Randomized	0.513
Stroke	Random Forest	none	0.506
Kidney	Random Forest	none	0.481
Cancer	Random Forest	Randomized	0.455

Analysis

None of the models are good predictors (maximum R^2 value of 0.568)
The tree ensemble methods work better than the SVM or OLS

Health outcomes data

- Obtain objective health data

Environmental pollution over time

- Include only people who have been living in   the same census tract for 10 years
- Use more than one year of environmental data