Predicting Health Outcomes from Environmental Pollutants

Predicting

Asthma

Cancer

COPD

Congestive Heart Failure

Kidney Disease

Stroke

From

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)

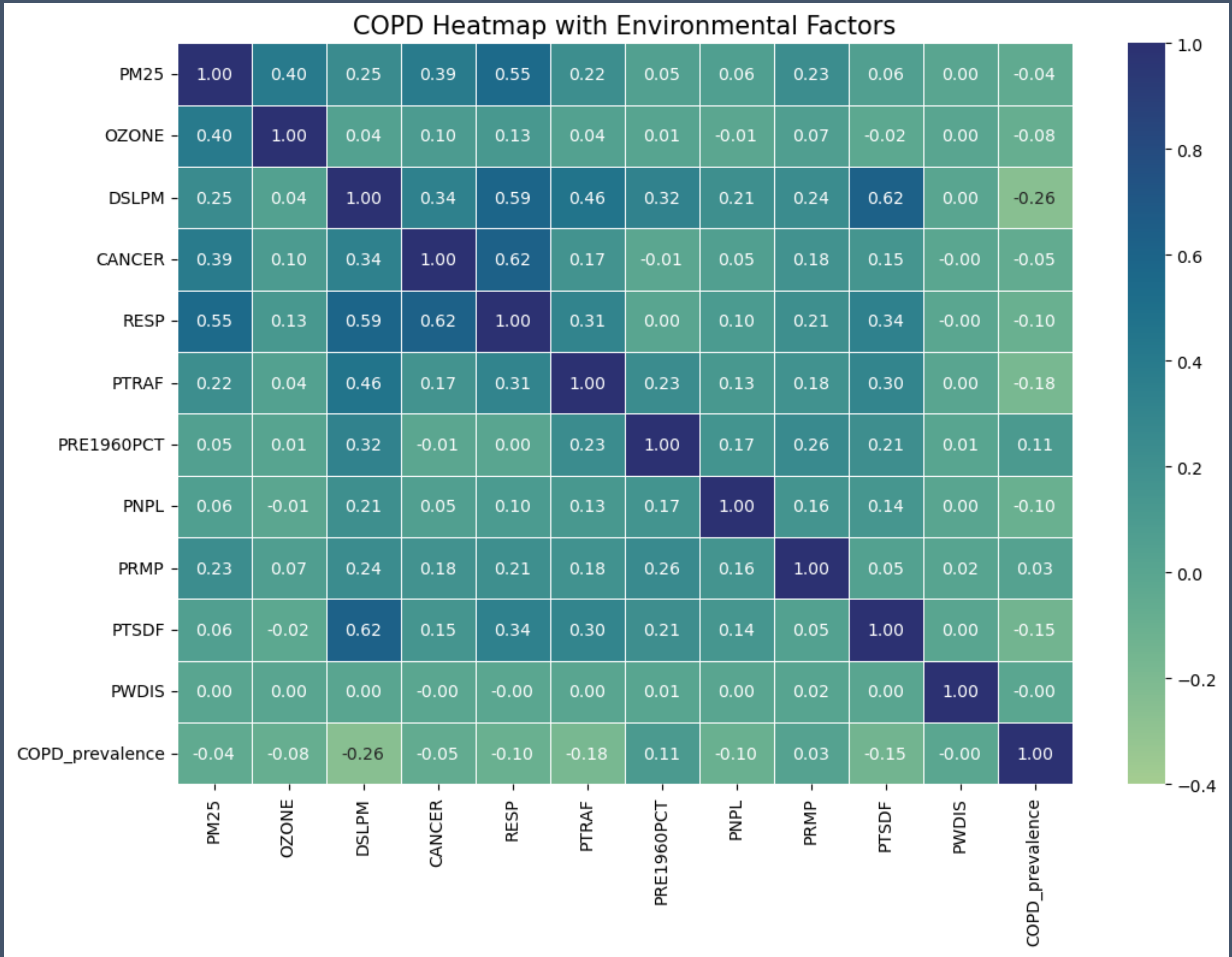
Proximity to National Priorities List (NPL) [superfund] sites

Proximity to Risk Management Plan (RMP) facilities

Proximity to Treatment Storage and Disposal facilities

Indicator for major direct dischargers to water

Exploratory Data Analysis

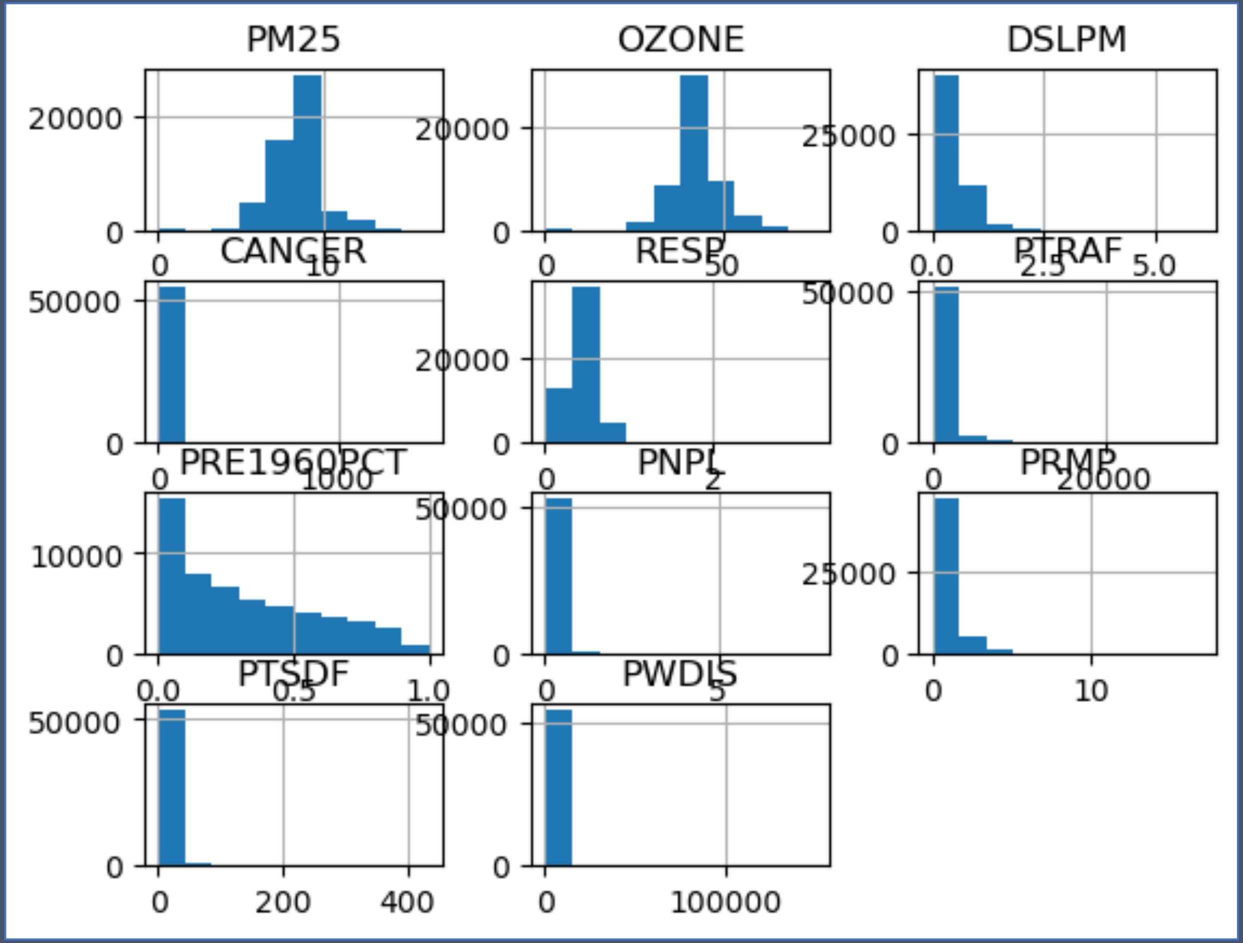


COPD Heatmap with Environmental Factors

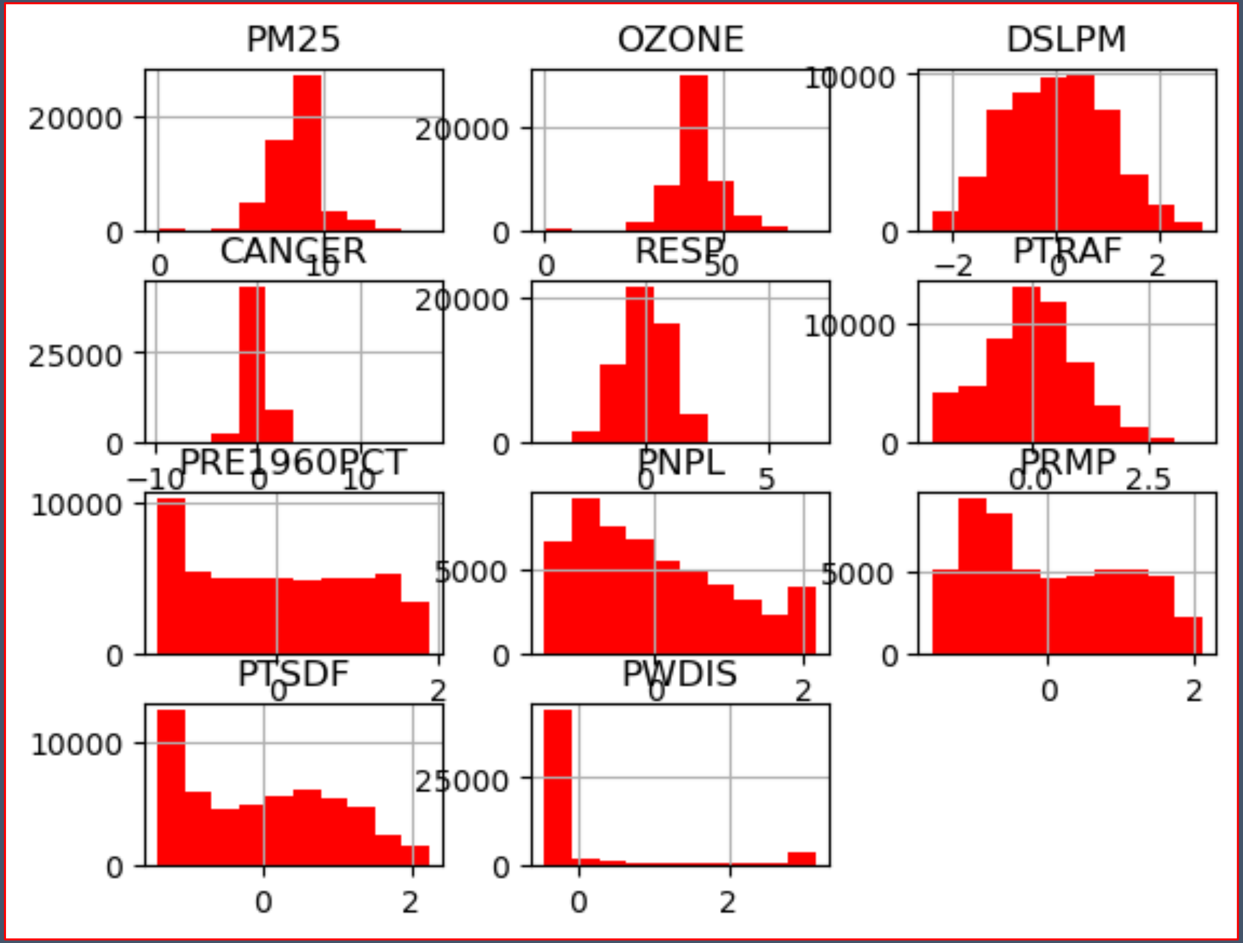
	PM25	OZONE	DSLPM	CANCER	RESP	PTRAF	PRE1960PCT	PNPL	PRMP	PTSDF	PWDIS	COPD_prevalence
PM25	1.00	0.40	0.25	0.39	0.55	0.22	0.05	0.06	0.23	0.06	0.00	-0.04
OZONE	0.40	1.00	0.04	0.10	0.13	0.04	0.01	-0.01	0.07	-0.02	0.00	-0.08
DSLPM	0.25	0.04	1.00	0.34	0.59	0.46	0.32	0.21	0.24	0.62	0.00	-0.26
CANCER	0.39	0.10	0.34	1.00	0.62	0.17	-0.01	0.05	0.18	0.15	-0.00	-0.05
RESP	0.55	0.13	0.59	0.62	1.00	0.31	0.00	0.10	0.21	0.34	-0.00	-0.10
PTRAF	0.22	0.04	0.46	0.17	0.31	1.00	0.23	0.13	0.18	0.30	0.00	-0.18
PRE1960PCT	0.05	0.01	0.32	-0.01	0.00	0.23	1.00	0.17	0.26	0.21	0.01	0.11
PNPL	0.06	-0.01	0.21	0.05	0.10	0.13	0.17	1.00	0.16	0.14	0.00	-0.10
PRMP	0.23	0.07	0.24	0.18	0.21	0.18	0.26	0.16	1.00	0.05	0.02	0.03
PTSDF	0.06	-0.02	0.62	0.15	0.34	0.30	0.21	0.14	0.05	1.00	0.00	-0.15
PWDIS	0.00	0.00	0.00	-0.00	-0.00	0.00	0.01	0.00	0.02	0.00	1.00	-0.00
COPD_prevalence	-0.04	-0.08	-0.26	-0.05	-0.10	-0.18	0.11	-0.10	0.03	-0.15	-0.00	1.00

Normalizing Data

Before



After



Types of Regression Modeling Done

Multiple OLS

Random Forest

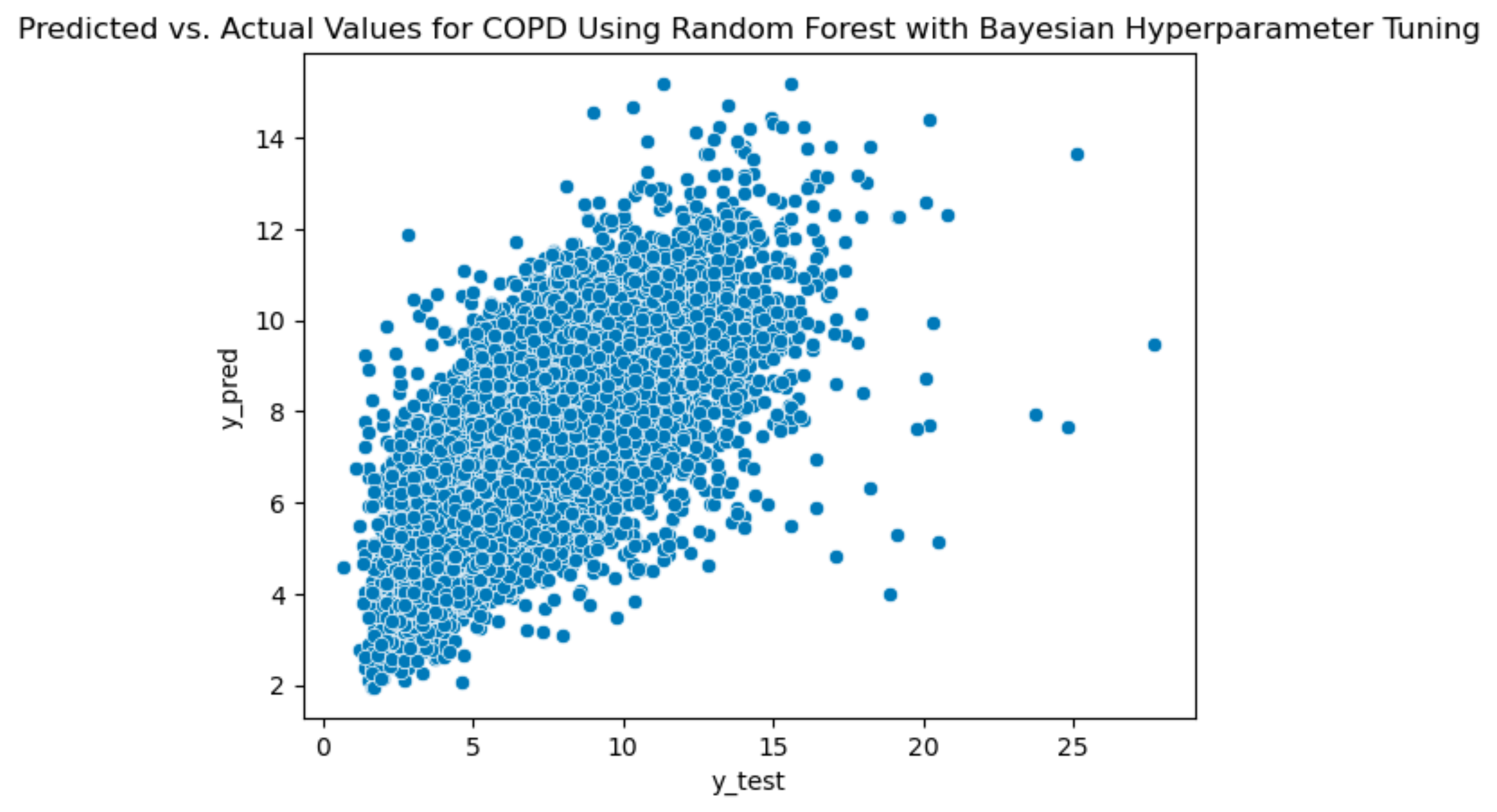
AdaBoost

Gradient Boost

XGBoost

LightGBM

SVM



Predicted vs. Actual Values for COPD Using Random Forest with Bayesian 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

Difficulties

Health outcomes data

Environmental pollution over time

Future Work

Obtain objective health data

Include only people who have been living in the same census tract for 10 years

Use more than one year of environmental data.