Air Quality Analysis - EPA

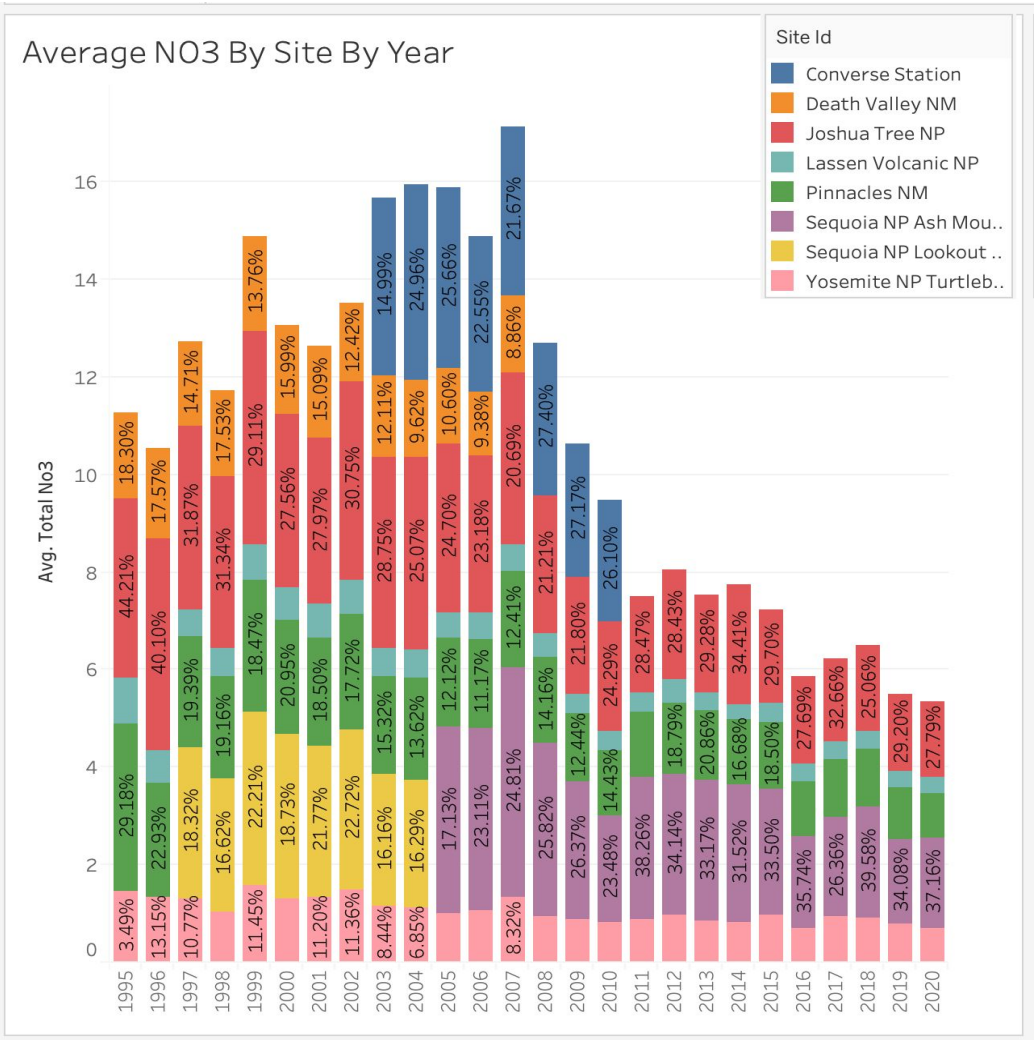
DataScienceGO Hackathon
Team 4

Observations

TSO4	0.702431
TNO3	0.775245
TNH4	0.796080
Ca	0.564870
Mg	0.432192
Na	0.235053
K	0.522962
Cl	−0.068791
NSO4	0.307510
NHNO3	0.854559
WSO2	0.513682
TOTAL_SO2	0.509987
TOTAL_NO3	1.000000
FLOW_VOLUME	−0.073946
VALID_HOURS	−0.000851
STD2LOCAL_CF	0.010595
QA_CODE	0.022223
year	−0.199742

- Strong correlation between concentration of other pollutants
  with total NO3:
  - TOTAL So2
  - NH4
- Slightly negative correlation total NO3 with years → overall decline over time
- Summed the data per pollutant to study trends by site over time

		TOTAL_NO3	TOTAL_SO2	TNH4
SITE_ID	year			
PIN414	2007	101.1324	18.7562	12.0389
	2001	115.5411	19.6126	13.1032
JOT403	1999	209.2936	26.9765	24.1577
	1997	187.5133	27.2324	21.1358
YOS404	2017	25.4788	9.9396	7.4635
DEV412	1995	68.7100	13.0981	11.6989
LAV410	2019	9.1859	4.2158	3.1739
PIN414	2013	61.6263	12.5996	8.5730
DEV412	2006	67.8605	17.3251	15.2490
YOS404	2009	45.2119	14.5679	12.7276



Observations:

- Certain sites had missing data for Total NO3, i.e. Converse Station, Death Valley, appears that these sites stopped testing
- Sequoia NP - data taken from two separate locations over time - should be viewed continuously
- Beginning in 2012, each site saw a decreasing trend of NO3 resulting in Total NO3 decreasing over time to 2020

Predictions

Preprocessing:

- Filled chemicals missing data w/median at site

Model

- Feature selection: all numeric columns + one hot encoded object columns
- Feature engineering: monthly and weekly time periods
- Algorithms
    - Linear Regression: Sum of Squared Errors = 2.2e-06