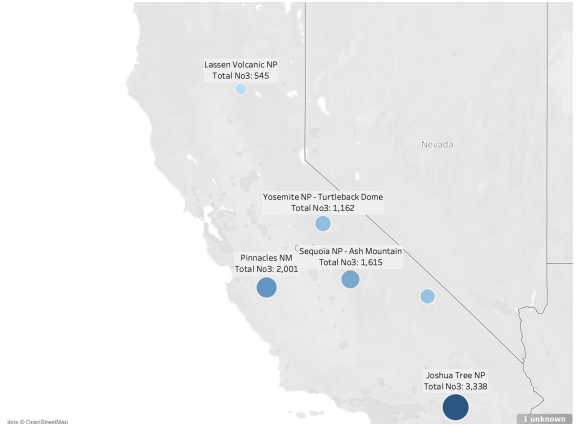
Predicting Total Nitrate (NO_3) in the Atmosphere

TEAM 2
Ali Khan
Marianna Carini
Sasha Prokhorova
Marquise Piton

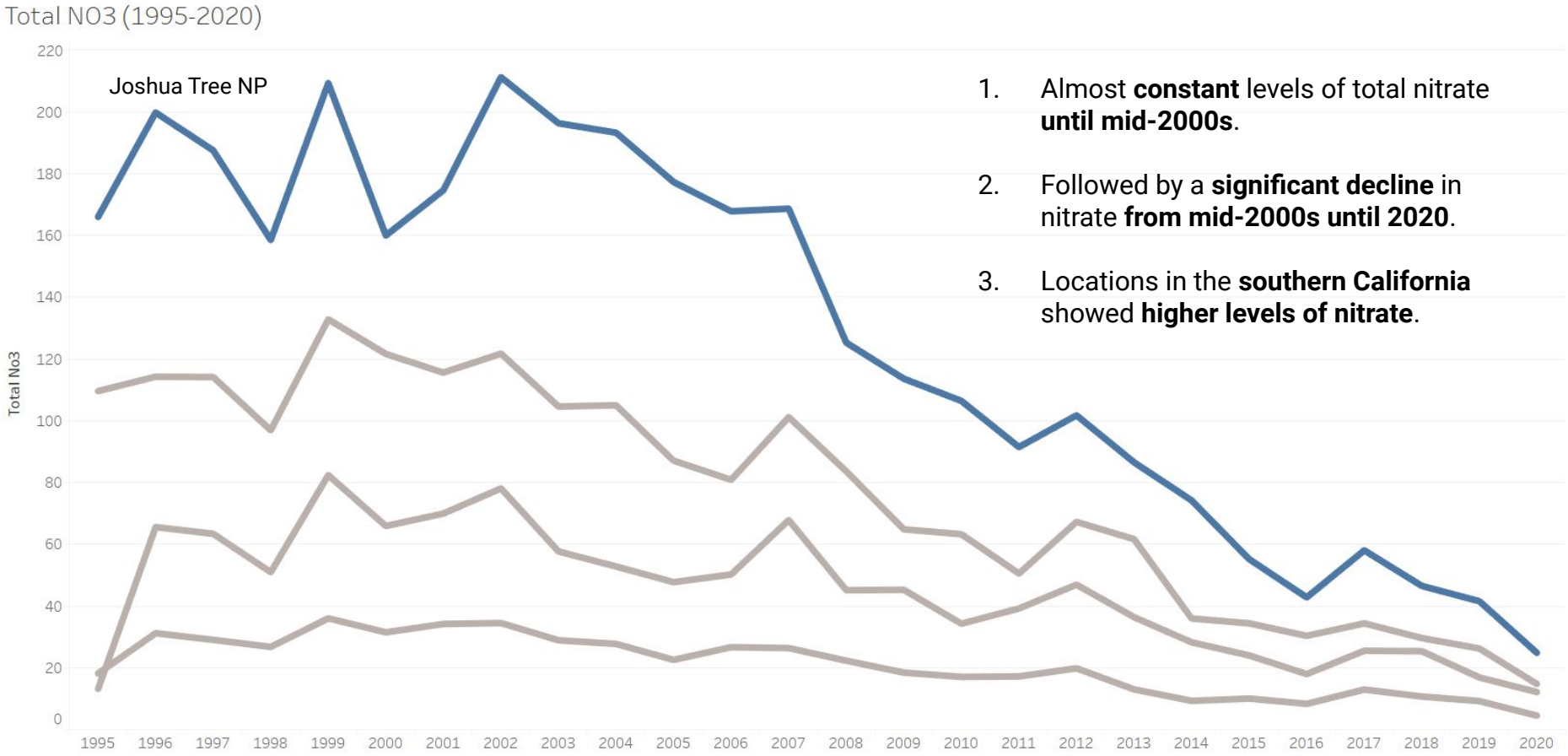
Protecting our ecosystems starts with understanding the fundamental threats to them.

In this project, we explore common environmental pollutants and build a predictive model to estimate the total nitrate (NO_3) in the atmosphere.

Exploratory Data Analysis



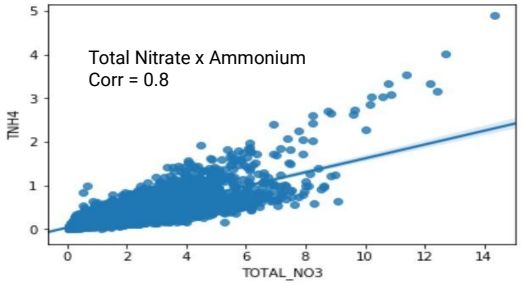
Total NO3 (ug/m3)

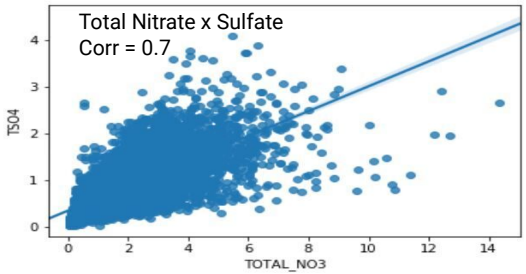


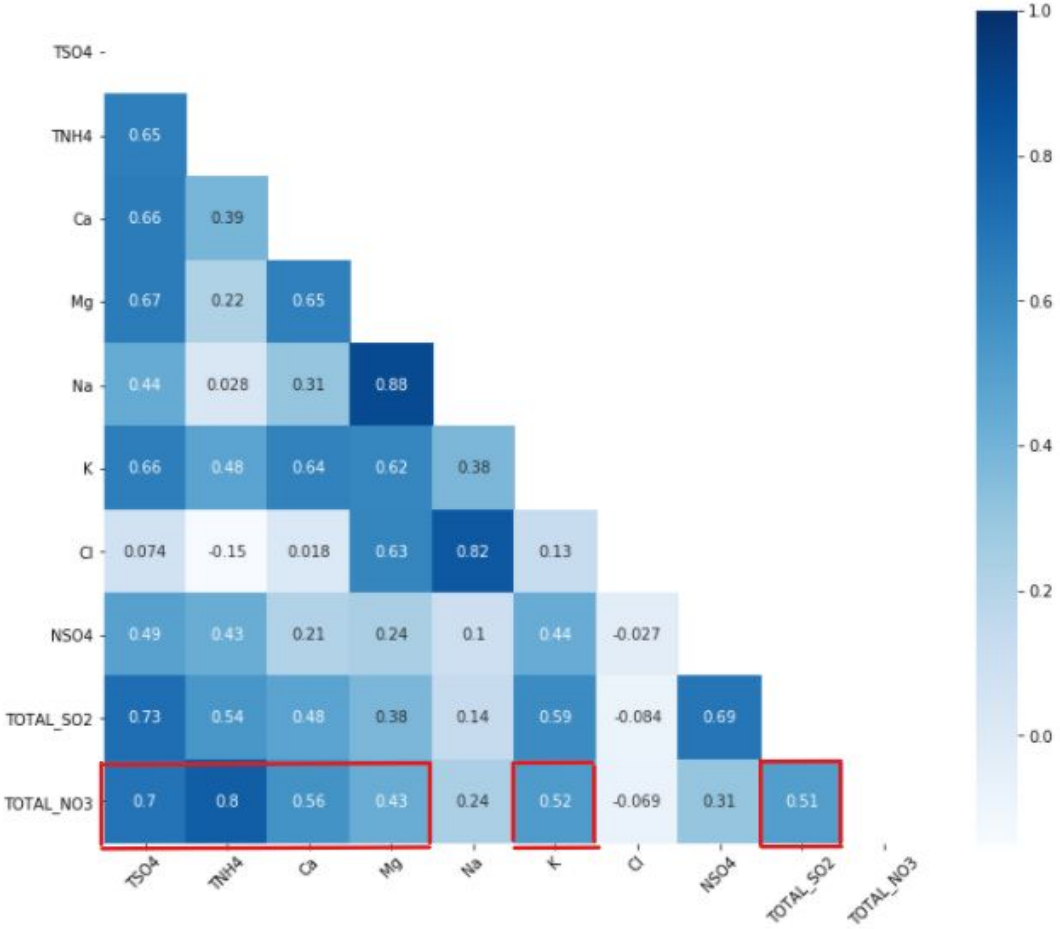
1. Almost **constant** levels of total nitrate **until mid-2000s**.
2. Followed by a **significant decline** in nitrate **from mid-2000s until 2020**.
3. Locations in the **southern California** showed **higher levels of nitrate**.

Correlation of Total Nitrate (in descending order):

1. **Ammonium (TNH4) 0.8**
2. **Sulfate (TSO4) 0.7**
3. Calcium (Ca) 0.56
4. Potassium (K) 0.52
5. Sulfur Dioxide (SO2) 0.51
6. Magnesium (Mg) 0.43

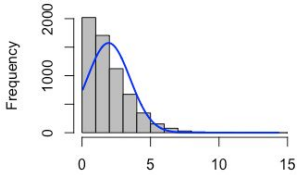






Modeling Total NO3

- Using the variables of interest fromt the EDA we were able to move to our modeling
- Because the NO3 data was right skewed, we decided to use a Gamma model
- Below is a summary of the model




```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.10972    0.02024   5.421 6.20e-08 ***
TNH4         3.13643    0.04582  68.457  < 2e-16 ***
TSO4         0.48518    0.03610  13.441  < 2e-16 ***
Ca           2.32033    0.10302  22.522  < 2e-16 ***
K           -0.67947    0.32416  -2.096   0.0361 *
WSO2        -0.23592    0.04240  -5.565 2.77e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7825 on 4926 degrees of freedom
  (1227 observations deleted due to missingness)
Multiple R-squared:  0.7475,    Adjusted R-squared:  0.7472
F-statistic:  2916 on 5 and 4926 DF,  p-value: < 2.2e-16
```

75% R-squared value

Variables are statistically significant

88% R-squared value

OLS Regression Results

Dep. Variable:	TOTAL_NO3	**R-squared (uncentered):**	0.878
Model:	OLS	**Adj. R-squared (uncentered):**	0.878
Method:	Least Squares	**F-statistic:**	1.476e+04
Date:	Sat, 24 Oct 2020	**Prob (F-statistic):**	0.00
Time:	14:19:33	**Log-Likelihood:**	-7887.8
No. Observations:	6159	**AIC:**	1.578e+04
Df Residuals:	6156	**BIC:**	1.580e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
TSO4	1.0898	0.027	40.838	0.000	1.037	1.142
TNH4	3.0416	0.047	64.486	0.000	2.949	3.134
WSO2	-0.2951	0.042	-7.056	0.000	-0.377	-0.213

Omnibus:	624.858	**Durbin-Watson:**	0.654
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1878.262
Skew:	0.537	**Prob(JB):**	0.00
Kurtosis:	5.483	**Cond. No.**	5.58

Appendix

Limitations and Assumptions

Limitations

1. Data limited to only California
2. Some locations have different periods of data
3. Missing measurements for some variables (Ca, Mg, K, Na, Cl)

Assumptions:

1. Correlation independent of time
2. No significant climate changes in California since 1995
3. No effect of topography on the difference between locations

Conclusions

- Variables of interest for Total NO3: TSO4, TNH4, WSO2, Ca, K
- Variables of interest for Total NO3: TSO4, TNH4, WSO2