Time Series Visualization of USA Pollution Data

By: Yash Gokhale, CMU
Data Science for Technology, Innovation and Public Policy

Background:

Tracking and modeling emissions is an important operation, so as to ensure that the spatial and temporal variance in these parameter values is identified.

Aim:

The project was aimed at able to convert the large scale emission data available into a user-friendly, easily decipherable content.

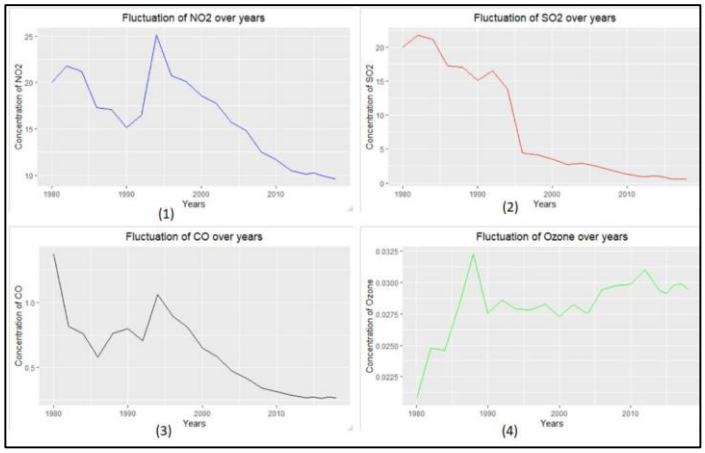
Tools Used:

Data source: [US EPA Air Pollution Data]

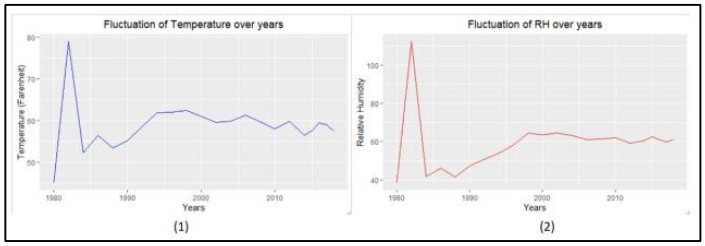
Languages used: Python, R

Libraries used: dplyr, ggplot2, tidyverse, lubridate, numpy, pandas, matplotlib, tkinter

Time Variance of Pollutants:



Variation of independent variables over time



Variation of dependent variables over time

Models Tried:

- *Linear Regression Model*
- *Level Log Model*
- *Log Log Model*

Number of dependent variables:

4 pollutant concentrations

50 States in USA

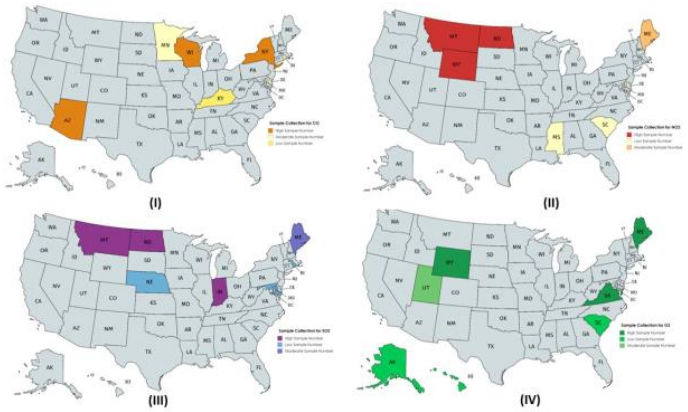
Time points from 1980

Model	Temperature	%RH
Linear Regression Model	0.8149	0.5574
Level Log Model	0.6916	0.506
Log Log Model	0.6823	0.4921

Average Test set accuracy for 4 cross validation

Sample Collection Methodology

The number of samples to be collected for every state was identified and highlighted on the USA map for better visualization.



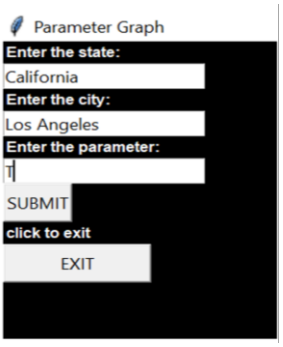
Spatial Distribution Graph for Sample Collections: I-CO, II-NO_2, III-SO_2, IV-O_3

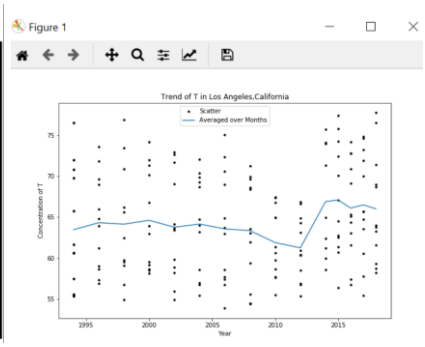
From the analysis, it was established that the number of samples to be collected varied greatly with the state in question. The entire ranking of pollutants for every state can be found at:

[Model Results]

Development of GUI using Python

A standalone GUI was developed using Tkinter to help visualize the pollutants over time, over a particular state and city.





The entire medium article can be found at: [Medium Article] and Github Repo at: [GitHub Code Repository]