|  |  |  |  |
| --- | --- | --- | --- |
| Machine Learning Model | Data Sets | Locations | Cleaned for Training |
| GaussianProcessRegressor  [Link](https://scikit-learn.org/stable/modules/generated/sklearn.gaussian_process.GaussianProcessRegressor.html?highlight=gauss#examples-using-sklearn-gaussian-process-gaussianprocessregressor)  [Model Tuning noted] | Kaggle Data Sets up to 2016  [for all 6 pollutants]  Wake-History.csv  Seattle-History.csv  NewYork-History.csv  Laramie-History.csv  LA-History.csv  ELPaso.csv | New York  Los Angeles  Raleigh  El Paso area  Laramie  Seattle | 6 files per location |
| GaussianProcessRegressor  [Link](https://scikit-learn.org/stable/modules/generated/sklearn.gaussian_process.GaussianProcessRegressor.html?highlight=gauss#examples-using-sklearn-gaussian-process-gaussianprocessregressor)  [Model Tuning noted] | EPA Daily Data – by CBSA  [for PM10, Ozone, NO2]  Files in Daily Data EPA | New York  Los Angeles  Raleigh  El Paso area  Laramie  Seattle | 3 files per location |
| GaussianProcessRegressor  [Link](https://scikit-learn.org/stable/modules/generated/sklearn.gaussian_process.GaussianProcessRegressor.html?highlight=gauss#examples-using-sklearn-gaussian-process-gaussianprocessregressor)  [Model Tuning noted] | Append EPA data to Kaggle data for PM10, Ozone, NO2]  Last set for predictions | New York  Los Angeles  Raleigh  El Paso area  Laramie  Seattle |  |

Data Set format:

“year”,” month”,”day”,”weight”,”flag”,”pollutant-name”

1958,03,29,4,0,MLO,316.10

1958,04,05,6,0,MLO,317.30

Notes:

Flag always = zero [0]

Weight is number of monitor readings

Date-Split out year, month, day

Station – use shortest code