|  |  |  |  |
| --- | --- | --- | --- |
| 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”,”station-name”,”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

Text book [www.gaussianprocess.org/gpm/chapters](http://www.gaussianprocess.org/gpm/chapters)

From Kaggle data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SITE ID | Units | Weight Measured | Monitoring Sites XXREF | County | State |
| SPNG | O3 1st Max Hour ppm | O3 1st Max Hour | 3801 Spring Forest RD | Wake | NC |
|  |  |  |  |  |  |

From EPA Daily Data – Final Sets XREF

Data Format and column order:

Target [DAILY\_AQI\_VALUE],

POLLUTANT [I.E. Daily Max 8-hour Ozone Concentration],

WEIGHT[DAILY\_OBS\_COUNT],

year,

month,

day,

station-name [Site Name],

Lat [SITE\_LATITUDE],

Long [SITE\_LONGITUDE]

NOTE: TO EXTRACT OUT YEAR, MONTH AND DAY FROM DATE: =TEXT(a2,” yyyy”); =TEXT(a2,”mm”); TEXT(a2,”dd”)

FOR each year set

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SITE Name | Units | Weight | Monitoring Sites XXREF | County | State |
| 2 NAMEs | Daily Max 8-Hour Ozone Concentration - ppm | 17 max per day | Map to Longitude and Latitude | Laramie | WY |
| 1 NAMED | Daily Max 8-Hour CO Concentration - ppm | 24 max per day |  |  |  |
| 2 Named | Daily Max 1-hour NO2 Concentration - ppb | 24 max per day |  |  |  |
| 3 Named | Daily Mean PM10 Concentration – ug/m3 SC | 1 per day |  |  |  |
| 2 Named | Daily Max 1-hour SO2 Concentration – ppb | 22 max per day |  |  |  |
| 3 Named | Daily Max 8-hour CO Concentration – ppm | 24 max per day | Map to Longitude and Latitude | El Paso | TX |
| 3 Named | Daily Max 1-hour NO2 Concentration - ppb | BTWN 19 and 24 |  |  |  |
| 6 Named | Daily Max 8-Hour Ozone Concentration - ppm | 17 max per day |  |  |  |
| 3 Named | Daily Mean Pb Concentration - ug/m3 LC | 1 per day |  |  |  |
| 5 Named | Daily Mean PM10 Concentration – ug/m3 SC | 1 per day |  |  |  |
| 3 Named | Daily Max 1-hour SO2 Concentration – ppb | 24 max per day |  |  |  |
| 18 Named | Daily Max 8-Hour CO Concentration - ppm | 24 max per day | Map to Longitude and Latitude | El Paso | TX |
| 18 Named | Daily Max 1-hour NO2 Concentration - ppb | BTWN 12 and 25 |  |  |  |
| 18 Named | Daily Max 8-Hour Ozone Concentration - ppm | 17 max per day |  |  |  |
| 10 Named | Daily Mean Pb Concentration - ug/m3 LC | 1 per day |  |  |  |
| 10 Named | Daily Mean PM10 Concentration – ug/m3 SC | 1 per day |  |  |  |
| 4 Named | Daily Max 1-hour SO2 Concentration – ppb | 24 per day |  |  |  |
| 8 named | Daily Max 8-Hour CO Concentration - ppm | 24 max per day | Map to Longitude and Latitude | NY | NY |
| 11 Named | Daily Max 1-hour NO2 Concentration - ppb | 24 max per day |  |  |  |
| 22 Named | Daily Max 8-Hour Ozone Concentration - ppm | 17 max per day |  |  |  |
| 3 Named | Daily Mean Pb Concentration - ug/m3 LC | 1 per day |  |  |  |
| 2 Named | Daily Mean PM10 Concentration – ug/m3 SC | 1 per day |  |  |  |
| 13 Named | Daily Max 1-hour SO2 Concentration – ppb | Up to 24 per day |  |  |  |
| 2 Named | Daily Max 8-Hour CO Concentration - ppm | Up to 24 per day | Map to Longitude and Latitude | Raleigh | NC |
| 2 Named | Daily Max 1-hour NO2 Concentration - ppb | Up to 24 per day |  |  |  |
| 2 Named | Daily Max 8-Hour Ozone Concentration - ppm | 17 per day |  |  |  |
| 1 Named | Daily Mean PM10 Concentration – ug/m3 SC | 1 per day |  |  |  |
| 1 Named | Daily Max 1-hour SO2 Concentration – ppb | 23 per day |  |  |  |
| 2 Named | Daily Max 8-Hour CO Concentration - ppm | Up to 24 per day | Map to Longitude and Latitude | Seattle | WA |
| 3 – Named | Daily Max 1-hour NO2 Concentration - ppb | Up to 24 per day |  |  |  |
| 5 – Named | Daily Max 8-Hour Ozone Concentration - ppm | 17 per day |  |  |  |
| 1 – Named | Daily Mean PM10 Concentration – ug/m3 SC | 1 per day |  |  |  |
| 1 – Named | Daily Max 1-hour SO2 Concentration – ppb | Up to 24 Per day |  |  |  |