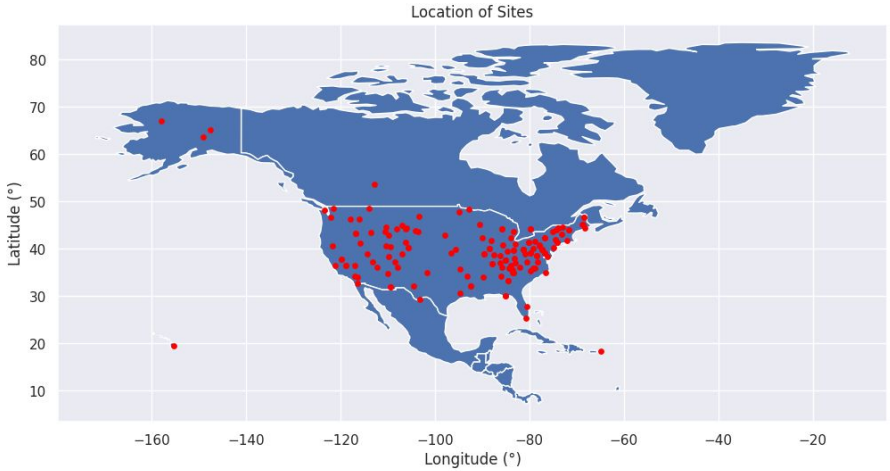
# Combating Air Pollution: Analyzing and Predicting Ozone Concentration

Guanyu Huo (gh2646), Duru Kahyaoglu (dk2565), Andrea Lopez (apl2171),
Daniel Pak (dcp2149), Naishadh Parmar (nnp2118)

# How Ozone Affects the Environment

- Ozone in the atmosphere is beneficial as it absorbs some radiation from the sun.
- However, ozone at lower elevations can harm the respiratory system of humans and other living organisms.
- Ozone is also an indicator of air pollution.
- Clean Air Status and Trends Network (CASTNET[1]) is an atmospheric monitoring program located across North America that records and publishes measurements of ozone concentrations and other features related to air quality and the atmosphere.
- Using the CASTNET dataset, we can predict the concentration of ozone as a regression style problem using the supervised learning framework.
- By forecasting ozone levels, this can lead to changes in public policy and actions to combat air pollution level.

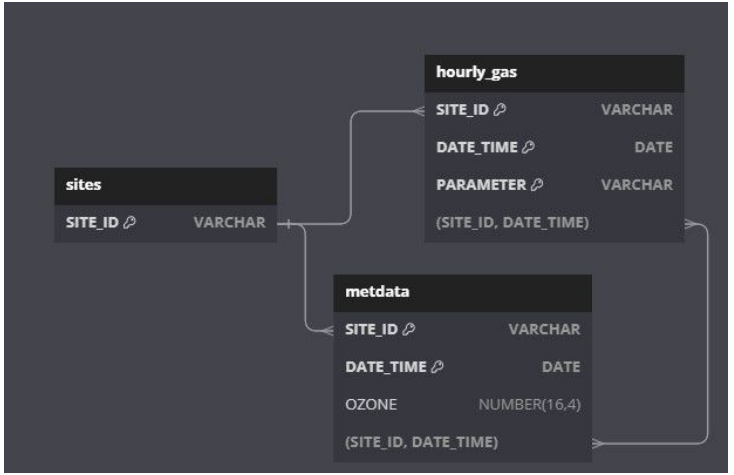



Location of Sites Where Atmospheric Measurements Are Taken

[1] https://www3.epa.gov/castnet/docs/CASTNET-Factsheet-2021.pdf

# Creating Preliminary Dataset

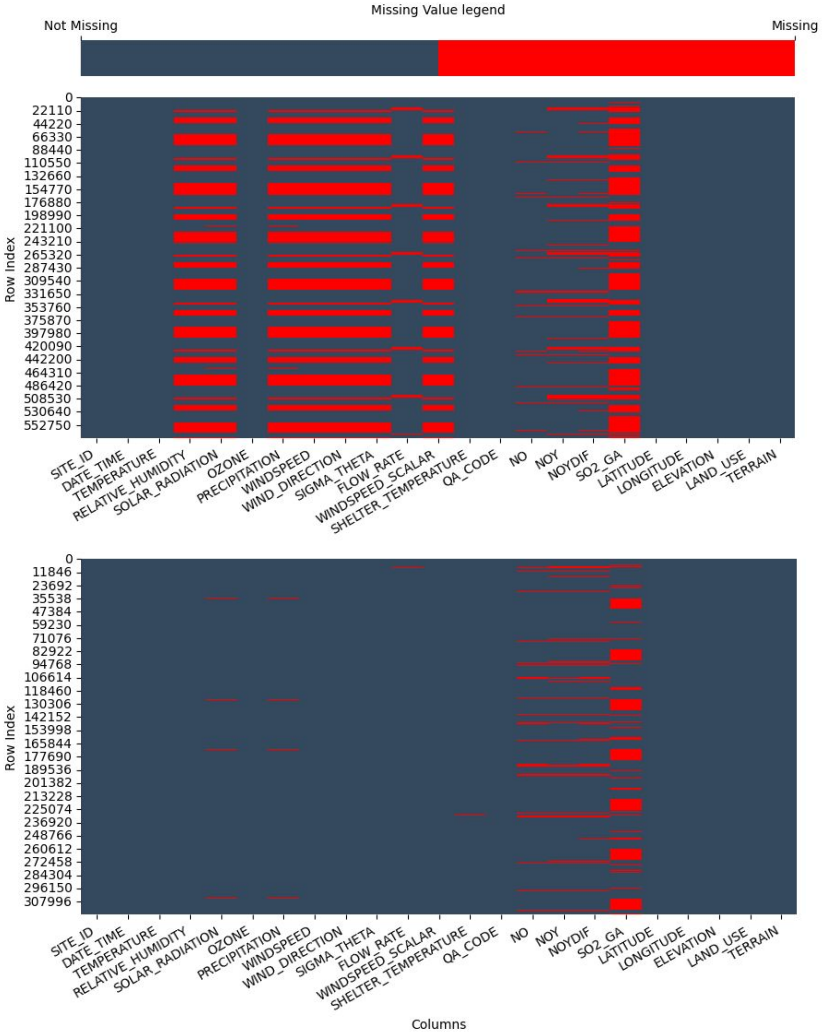
- We combined 3 datasets from CASTNET between the years 2013 to 2022: site, hourly gas, and meteorological data
- Specific columns were selected from the site dataset that are of interest like elevation and type of terrain, but others were removed that had no logical relationship to ozone like what agency manages the site
    - There are 6 features
- Filters were applied on the hourly gas dataset for excluding different parameters for gas measurements and including records that are valid measurements and have a high quality assurance level
    - This dataset was then pivoted to create columns with each type of parameter like NO2 or NH3 with the reported measurements
    - There are 17 features
- Filters were applied on the meteorological dataset for ozone records that are valid measurements and have a high quality assurance level
    - There are 29 features and 1 label
- The datasets were then joined with respect to the meteorological dataset on the corresponding primary keys/foreign keys
    - Site was joined with meteorological on SITE_ID
    - Hourly gas was joined with meteorological on SITE_ID and DATE_TIME
    - In total, there are 49 distinct features, 1 label and 574851 samples in our combined dataset




Data Model for Selected Tables in CASTNET

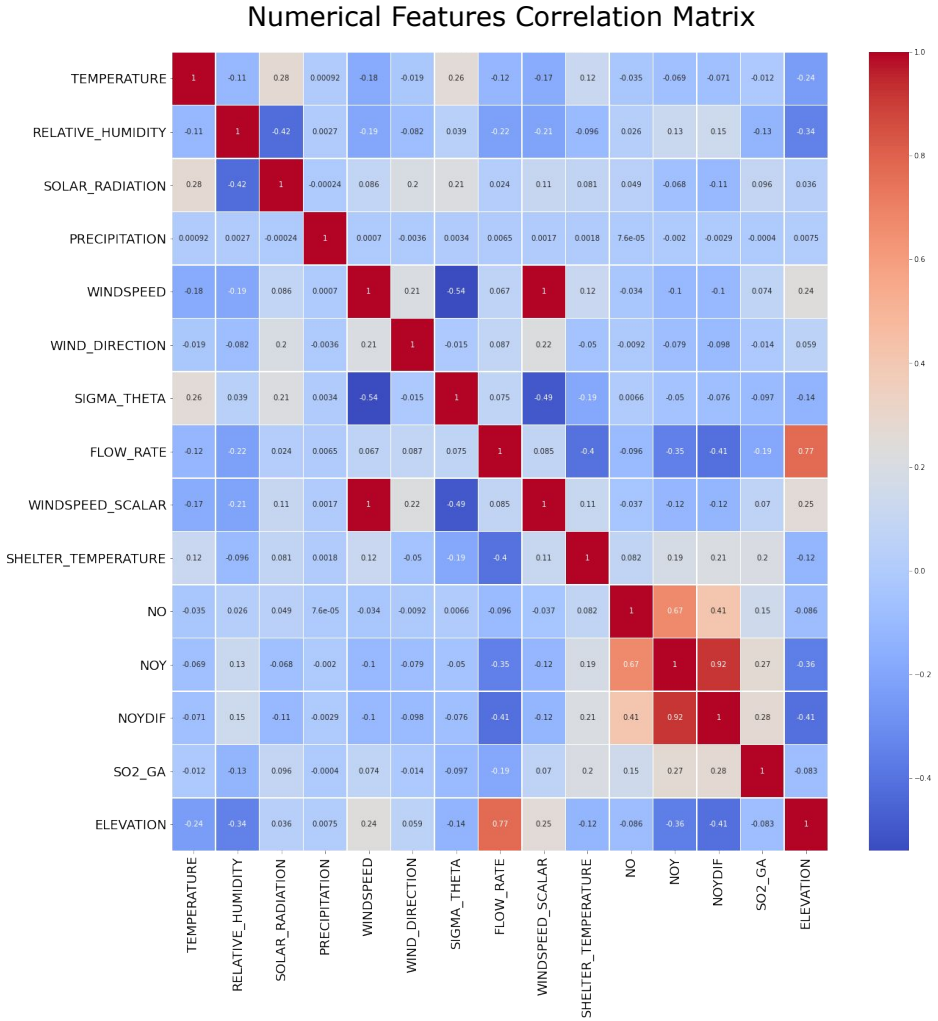
# Handling Missing Values

- Many of the features had missing values. We decided to drop all features with over 60% missing values.
- After that we observed that some sites have an entire feature missing, so we removed those sites entirely. This led to the transformation from the top figure to the bottom figure in terms of missing values.
- With this we ended up dropping approximately half the samples, but the missing values in the remaining half are now manageable.
- Finally the missing values are imputed using mean of each column.
- We also removed the quality of data measurement features since they were only useful for data cleaning.
- **Our final dataset consists of 6 distinct sites with 319835 rows, 22 features, and 1 label**



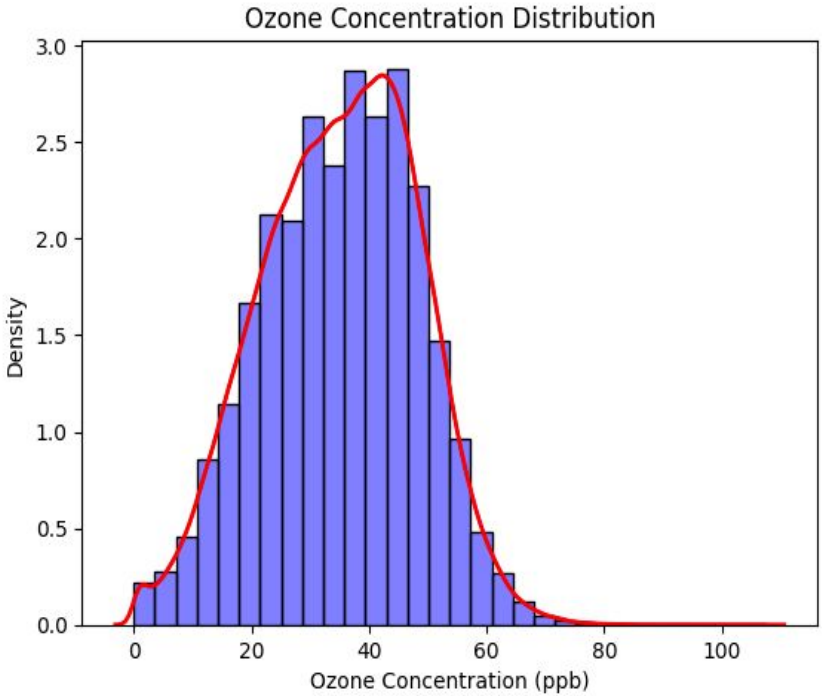
# Exploring Feature Correlation

- Excluding all non-numeric and categorical variables (like quality assurance, data quality of measurements, site ID, and land/terrain) we created a correlation matrix of the numerical variables.
- Wind speed and wind speed scalar are highly correlated because their meanings are direct functions of each other, so we decided to drop one of these features: windspeed.
- Sigma theta and wind speed scalar are also highly correlated because sigma theta is a function of wind direction, so we also decided to drop sigma theta.
- Other insights: Solar radiation and relative humidity are somewhat negatively correlated, which makes sense since more heat and sunlight can cause the air to be dry.



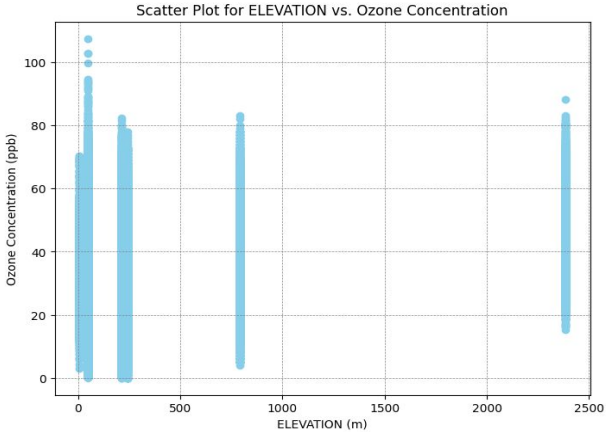
# Distribution of Ozone Concentration

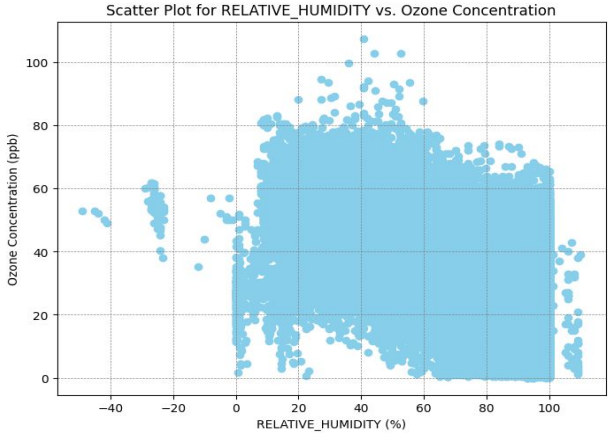
- Ozone concentration (in parts per billion) is the target variable of our model
- Therefore it requires a deeper understanding of how it's distributed to determine how to sample and split data into dev/validation/test datasets
- Distribution displays a slight right skew:
  - According to CASTNET, concentrations greater than 70 ppb can be unhealthy
  - There aren't many such records, in fact, less than 1% of overall data reflects this amount.
  - To ensure that the model covers these extreme cases, a sophisticated approach is required to split the dataset to ensure that higher ozone concentrations are included in the dev dataset.
  - Label transformation can also be done by taking the square root of the ozone concentration.

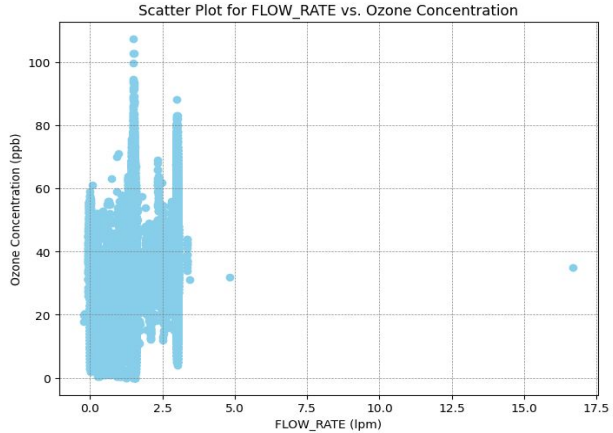


# Initial Observations on How Features Are Related to Ozone Concentration

- The top three features (elevation, relative humidity, and flow rate) that were most highly correlated with ozone concentration were plotted to get a sense of the relationship between measured environmental factors and ozone concentration.
- Examining these three plots, we can notice a few challenges with our features data that our ML model should address:
    - The cardinality of elevation is dependent on the number of sites and therefore we have a nonuniform distribution of elevation values. This is because most of the sites were removed while cleaning the data and we were left with only 6 sites and hence we only have 6 distinct elevation values.
    - As for relative_humidity and flow_rate, we can observe some outlier cases where they are much lower or higher than most of the values.
- The top features display a more complex relationship that requires beyond simple linear regression



# Feature Engineering, Encoding, Scaling, and Data Splitting

- From the DATE_TIME column, we created two new features: month and year. We later used these for splitting the data and will also investigate their effect on the label.
- We one hot encoded our categorical variables, LAND_USE and TERRAIN, which describe the land where the site is located.
- After dropping non-numerical features such as Site ID and quality assurance code, one hot encoding, and feature engineering, we ended up with 25 features and 1 label.
- Because we are working with time series data, we employed **structured splitting** and split our data by year.

| | | |
|---|---|---|
| Training | 2013-2018 | 60% of all data |
| Validation | 2019-2020 | 20% of all data |
| Testing | 2021-2022 | 20% of all data |

- We scaled our data to balance the impact of all features. We chose StandardScaler to preserve the original distribution of the data.

```
df['DATE_TIME']
df['month'] = df['DATE_TIME'].dt.month
df['year'] = df['DATE_TIME'].dt.year
```

```
from sklearn.preprocessing import OneHotEncoder

categorical_columns = ['LAND_USE', 'TERRAIN']

for feature in categorical_columns:
    encoder = OneHotEncoder()
    encoded_data = encoder.fit_transform(df[[feature]])
    df[encoder.categories_[0]] = encoded_data.toarray()

df.drop(columns=categorical_columns, axis=1, inplace=True)
```

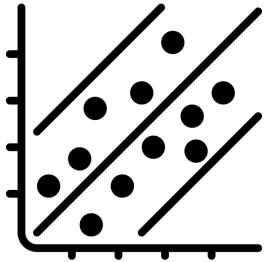
```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_val = scaler.transform(X_val)
X_test = scaler.transform(X_test)
```

# Proposed Machine Learning Techniques

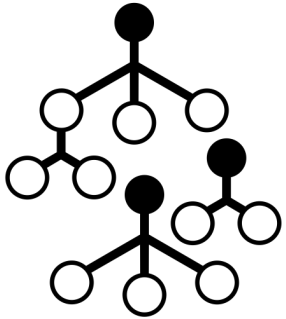
### Linear Regression Model

Linear Regression model could be used to analyze the linear relationships between Ozone concentration and various independent variables.



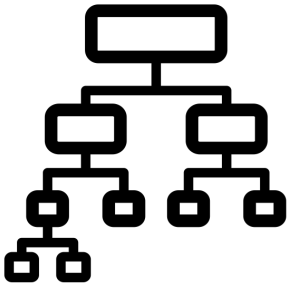
### Ensemble Methods

Ensemble methods could often yield better predictive performance by combining multiple models to reduce variance and bias. We would use ensemble models like Random Forests, Gradient Boosting Machines, XGBoost, etc.



### Decision Tree

Decision Tree could capture nonlinear relationships between different features and target Ozone concentration and provide an interpretable result.



# Baseline Model Performance

- As our problem is a regression task, $R^2$ score would explain the variance captured by the model and would be our main metric.
- Below are $R^2$ scores of our baseline models on the validation dataset, using the default parameters.

| Model | $R^2$ Score | Model | $R^2$ Score |
|---|---|---|---|
| Linear Regression | 0.568 | Decision Tree | 0.532 |
| Lasso Regression | 0.561 | Random Forest | 0.759 |
| Ridge Regression | 0.567 | XGBoost | 0.762 |

- Our baseline results are low, but employing hyperparameter tuning, model selection techniques, and using more robust models will lead to better performance.

# Next Steps

- **Model Selection**

  Based on data analysis and baseline models' evaluation, choose the most appropriate machine learning models for air pollution problem.

  We may experiment with splitting ozone concentrations into a few levels and treating our problem as a classification problem with models better suited for classification.

- **Hyperparameter Tuning**

  Employ Grid Search and Cross Validation to select and evaluate hyperparameters to improve model performance.

- **Model Interpretation**

  Explain output of model using feature importance, SHAP values and LIME to explore potential internal causal relationships within the data and offer guidance to reduce air pollution.