Recursive feature elimination with cross validation was used for feature selection, with the coefficient of determination as the scoring metric.

More work needed for RF

* Include location encoder (GEOID) into the model --> higher importance value == less differences between specific locations
* Include season encoder (month)?
* Train/test split (i.e. 70/30), evaluate performance on test set [DONE]
  + Evaluate with test set –
  + [[sklearn.ensemble.RandomForestRegressor.predict](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html#sklearn.ensemble.RandomForestRegressor.predict)](https://scikit-learn.org/stable/modules/model_evaluation.html#using-multiple-metric-evaluation)
  + [3.3. Metrics and scoring: quantifying the quality of predictions — scikit-learn 0.24.1 documentation (scikit-learn.org)](https://scikit-learn.org/stable/modules/model_evaluation.html#using-multiple-metric-evaluation)
* Calculate permutation importances
  + Use [sklearn.inspection.permutation\_importance](https://scikit-learn.org/stable/modules/generated/sklearn.inspection.permutation_importance.html), like in (1)
* Determine best-performing parameters/features using SciPy spearmanr for Spearman's Rank correlation – like in ([[1]](#footnote-1))
  + [scipy.stats.spearmanr](https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html)
* Cite relative risks (RRs)[[2]](#footnote-2) from PM2.5, NO2, Ozone (Table 2) ?
* Include more datasets [DONE]
  + Socioeconomic data[[3]](#footnote-3) - “USDA ERS data” folder
    - Yearly median income per county; simply use yearly value for the months in calendar year
  + GES DISC – “NASA GES DISC” folder – total column (dont use over ground ozone)
    - Ozone – daily, up-to-date data → implies predicting/modeling potential
    - NO2 maybe; much lower impact than ozone 2 , but concentrations are higher in USA/developed countries
  + EPA AQ data – ground ozone
    - Use AQI instead of ground ozone, since AQI data is more complete; easier to interpolation/imputation

4/22

1. ~~clean AQI data so that counties with incomplete interpolation are considered missing, so they are included in imputation~~
2. ~~Run CV for imputation strategies~~
3. ~~Run RFECV again, including AQI data from best imputation method~~
4. ~~Using the RFECV result, run GridSearchCV (or RFECV, if possible) with 1-2 month lag values included for the appropriate features~~
5. ~~Run SciPy's 'spearmanr' for Spearman's Rank correlation~~
6. Run Sklearn permutation importance
   * Follow same method in Cholera paper

Maybe run NN. Example: [GridSearchCV with MLPRegressor with Scikit learn - Data Science Stack Exchange](https://datascience.stackexchange.com/questions/52348/gridsearchcv-with-mlpregressor-with-scikit-learn)

* [sklearn.neural\_network.MLPRegressor — scikit-learn 0.24.1 documentation](https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html)
* [sklearn.preprocessing.MinMaxScaler — scikit-learn 0.24.1 documentation](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html)

1. [Cholera Risk: A Machine Learning Approach Applied to Essential Climate Variables (mdpi.com)](https://www.mdpi.com/1660-4601/17/24/9378/htm) [↑](#footnote-ref-1)
2. [Estimates of the Global Burden of Ambient PM2.5, Ozone, and NO2 on Asthma Incidence and Emergency Room Visits | Environmental Health Perspectives](https://ehp.niehs.nih.gov/doi/10.1289/EHP3766) [↑](#footnote-ref-2)
3. [USDA ERS - County-level Data Sets](https://www.ers.usda.gov/data-products/county-level-data-sets/) [↑](#footnote-ref-3)