Data Imported

Data Transformation

The data was transformed for analysis. The dataset from the CMC/CBP data download (https://drive.google.com/drive/folders/19HYUC5SLj7EV3Ui4fMhHBk_hnxcSY-PJ?usp=sharing) was filtered for total nitrogen. Then aggregated by date, longitude, latitude with a mean of the measure value - to obtain one reading for each day samples were taken.

Weather Data

Weather data was obtained from data provided by North American Regional Reanalysis - NARR (https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/north-american-regional-reanalysis-narr) . NARR is a regional reanalysis of North America containing temperatures, winds, moisture, soil data, and dozens of other parameters. The NARR model takes in, or assimilates, a great amount of observational data to produce a long-term picture of weather over North America. (from https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/north-american-regional-reanalysis-narr)

Features obtained from NARR from 1998 to 2020:
Downloaded from site: ftp://ftp.cdc.noaa.gov/Datasets/NARR/monolevel

- Air Temperature
- Humidity
- Cloud cover
- Surface Air Temperature
- Surface Runoff
- Wind components
- Precipitation

The relevant dates and location were extracted and joined with each pollutant observation & location (longitude/latitude).

Land Cover Data

Land Cover data was downloaded from the Multi-Resolution Land Characteristics Consortium (MRLC) National Land Cover Viewer. https://www.mrlc.gov/viewer/. A bounding box of the watershed (-80.44707800, 36.73004000, -74.83524400, 42.80672000) was used to download the relevant 2016 NLCD data, in the form of a .tiff raster file. The raster file displayed 20 different colors that corresponded with different types of land cover. Legend can be found here. https://nrcs.app.box.com/v/huc/file/532373547877.

This was used with the watershed HUC12 boundaries shapefile provided by the organizers of hack the bay :https://nrcs.app.box.com/v/huc/file/532373547877.

The two files were loaded into QGIS, watershed HUC12 boundaries shapefile was used to mask the land cover data and provide the land cover features of the tiff file by HUC12 code. The zonal statistics package was used on the resulting output to determine the mean pixel value of each HUC12 segment. Zonal histogram was also performed on each segment to count the totals for each pixel color/value. This layer was downloaded from QGIS and imported into python
In Python the layer features were imported and each HUC12 segment histogram was normalized so that each HUC12 segment had a total sum of 1. The information was then merged with the existing data on each HUC12 location.

The resulting data imported and joined with the existing data:

- Land cover % usage for each code in the nlcd legend https://www.mrlc.gov/data/legends/national-land-cover-database-2016-nlcd2016-legend
- The mean of the pixel values for each segment
- The area of each segment in acres

Features created

HUC_12_enc

This feature was created using mean target encoding for each HUC12 code. To address data leakage and introduce regularization a k-fold method was utilized, using the out of fold mean.
https://mlbook.explained.ai/catvars.html#target-encoding ,
https://www.geeksforgeeks.org/mean-encoding-machine-learning/
https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-munging/target-encoding.html

This feature is preferred over label encoding the HUC codes because there were more than 300 HUC12 locations, and One Hot Encoding because OHE would introduce a large amount of new features to the data. The mean encoding introduces the correlation between the categories and the target variable in one additional feature.

Distance from Outflow

Distance from each sample location (lat/long) from the outflow of the Bay (36.995833, -75.959444). The outflow location was determined from https://en.wikipedia.org/wiki/Chesapeake_Bay and distance measured as the geodesic distance between two coordinates, in miles.

Nitrate Oxide emitted from Point Sources by year

This feature represents all of NO2 pollutant in pounds from correlated point source locations in the airshed by year for each HUC12 segment.
This information was from the Air Emissions Dataset obtained from https://echo.epa.gov/tools/data-downloads . Combined air emissions data for stationary sources from four EPA air programs: National Emissions Inventory (NEI), Greenhouse Gas Reporting Program (GHGRP), Toxic Release Inventory (TRI), and Clean Air Markets. Emissions are presented as facility-level aggregates and organized by pollutant and EPA program.
The data was loaded and filtered to only states within the Chesapeake Bay Watershed (states determined from https://www.cbf.org/about-the-bay/maps/geography/the-chesapeakes-airshed.html and https://www.chesapeakebay.net/what/maps/chesapeake_bay_airshed#:~:text=The%20Airshed%20Model%20covers%20the,cell%20measures%20eighty%20kilometers%20square.)
The dataset was then filtered to only the Nitrogen Oxide pollutant emission. The original dataset was transformed and aggregated by HUC12 code and year and merged by year with the pollutant emissions of NO2 by year by Point Source. The data was correlated and filtered to only correlations >=.6 and <1. All correlated point sources for each HUC12 segment were aggregated by year. For those missing values - a mean of the HUC12 correlated point sources was used.