EY Data Challenge: Urban Heat Islands

Benjamin Nicholson and Jonah Zembower

EY Data Challenge: Urban Heat Islands

What? The EY Data Challenge has been operating since 2020 and is focused on the application of machine learning and statistical analysis to solve earth science problems

When? January 20 2025 - March 20 2025

Who? Applicants from all over the world (currently at 1200+ registrations)

Why? Guided experience through using satellite imagery APIs and implementing machine learning into the creation of a regression model

Urban Heat Islands Background Information

The Challenge

- Temperature data has been recorded at different locations across Manhattan and the Bronx in New York City
- This is used to formulate the Urban Heat Index (UHI)
- Goal: Predict the UHI for a given location.

Data Available

Training UHI Index Data

- Includes the longitude and latitude with respective UHI there
- Bronx and Manhattan data collected for air temperature and other variables

Satellites

- European Sentinel-2 Optical Satellite
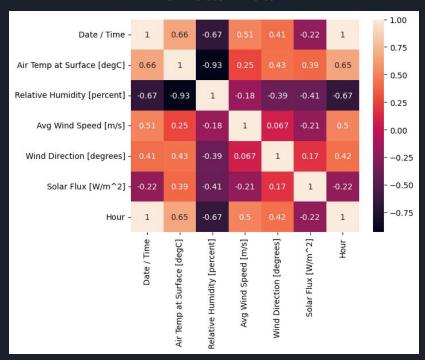- NASA Landsat Optical Satellite

Open Sourced New York Data

- Includes building height, roads, distance to shoreline, wind tunnels, tree canopy
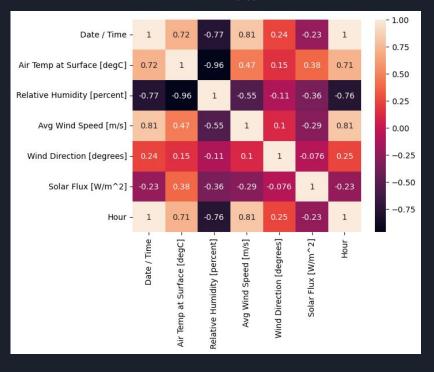
EDA

1. Looked at the correlation of the features.
2. Showcased Data Collected:
   a. UHI Index
   b. Air Temp at Surface [deg C]
   c. Relative Humidity [percent]
   d. Avg Wind Speed [m/s]
   e. Wind Direction [deg]
   f. Solar Flux [W/m^2]
3. Planetary Computer Data:
   a. Surface Temperature
   b. Median Composites
   c. NDVI
   d. NDBI
   e. NDWI
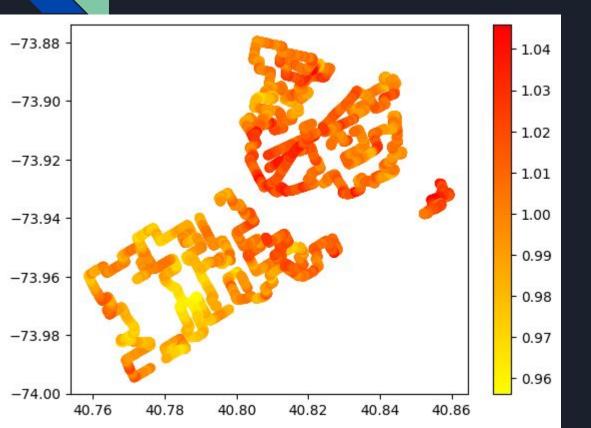4. Open Source New York Data

Correlation of Base Features
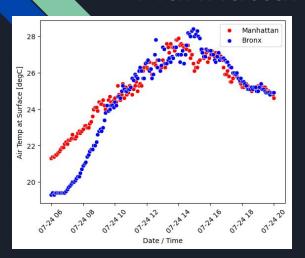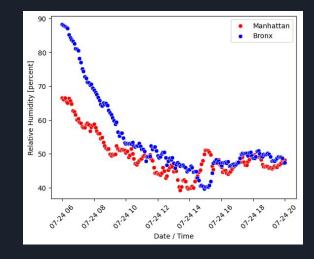
Manhattan Data

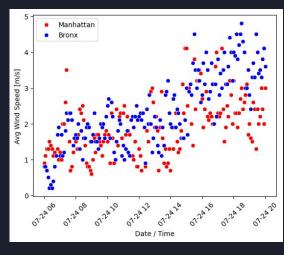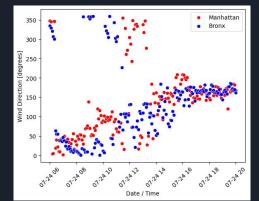
Bronx Data

Training UHI Index Values

- Includes longitude and latitude values
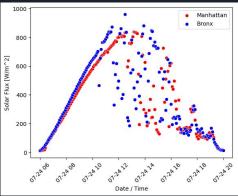- UHI index at each point
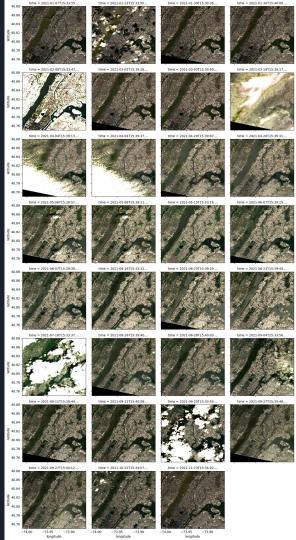
Manhattan Versus Bronx

- Air Temp at Surface [deg]
- Relative Humidity [percent]
- Avg Wind Speed [m/s]
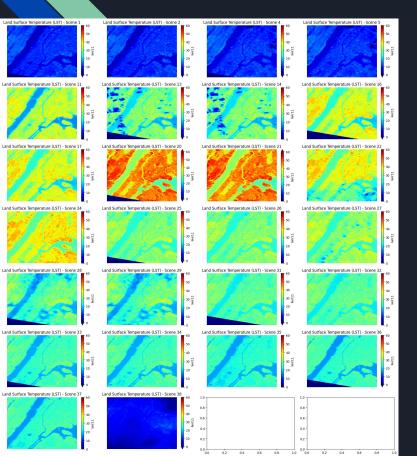- Wind Direction [degrees]
- Solar Flux [W/m^2]

Planetary Computer Scenes

- Images related to longitude and latitude selected from landsat satellite
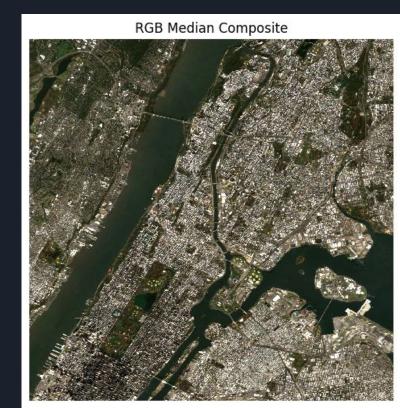- There were a total of 31 scenes of pictures in 2021

Planetary Computer Land Surface Temperature

- Land surface temperature related to the detected surface temperature in 2024
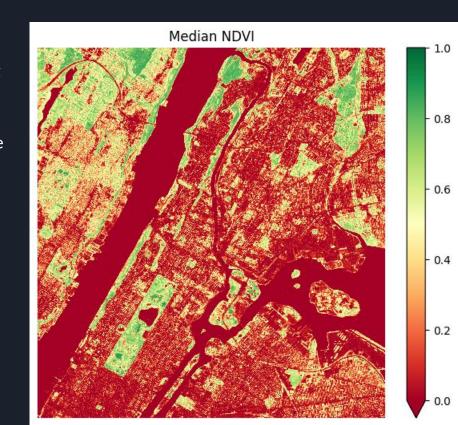- This was found using the Landsat Satellite

Planetary Median Composite

- Using normal xarray operations, we can compute the median pixel value at that point over many images.
- This chooses a best image across multiple captures in a given year or timeframe.
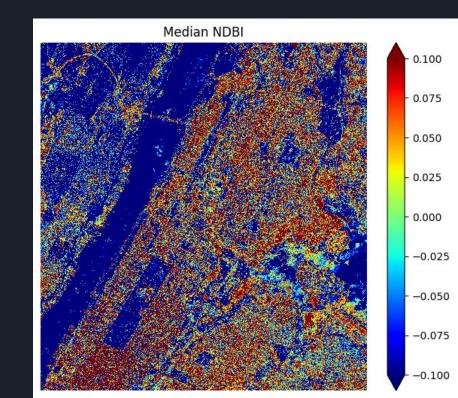

RGB Median Composite

Planetary Computer NDVI

- The NDVI is a calculated value describing the vegetation index gradient.
- Higher values indicate more vegetation.
- This is found using certain bands from the Sentinel-2 Satellite.

Median NDVI

Planetary Computer NDBI

- The NDBI is calculated as a gradient ratio for the certain bands from the Sentinel-2 Satellite.
- Showcases higher values reflecting more urbanization.

Median NDBI

Planetary Computer NDWI

- The NDWI is a calculated ratio using the Sentinel-2 Satellite bands.
- This highlights areas of water in blue and areas with little water in more red.

Median NDWI