EY Data Challenge: Urban Heat Islands

Benjamin Nicholson and Jonah Zembower

EY Data Challenge: Urban Heat Islands

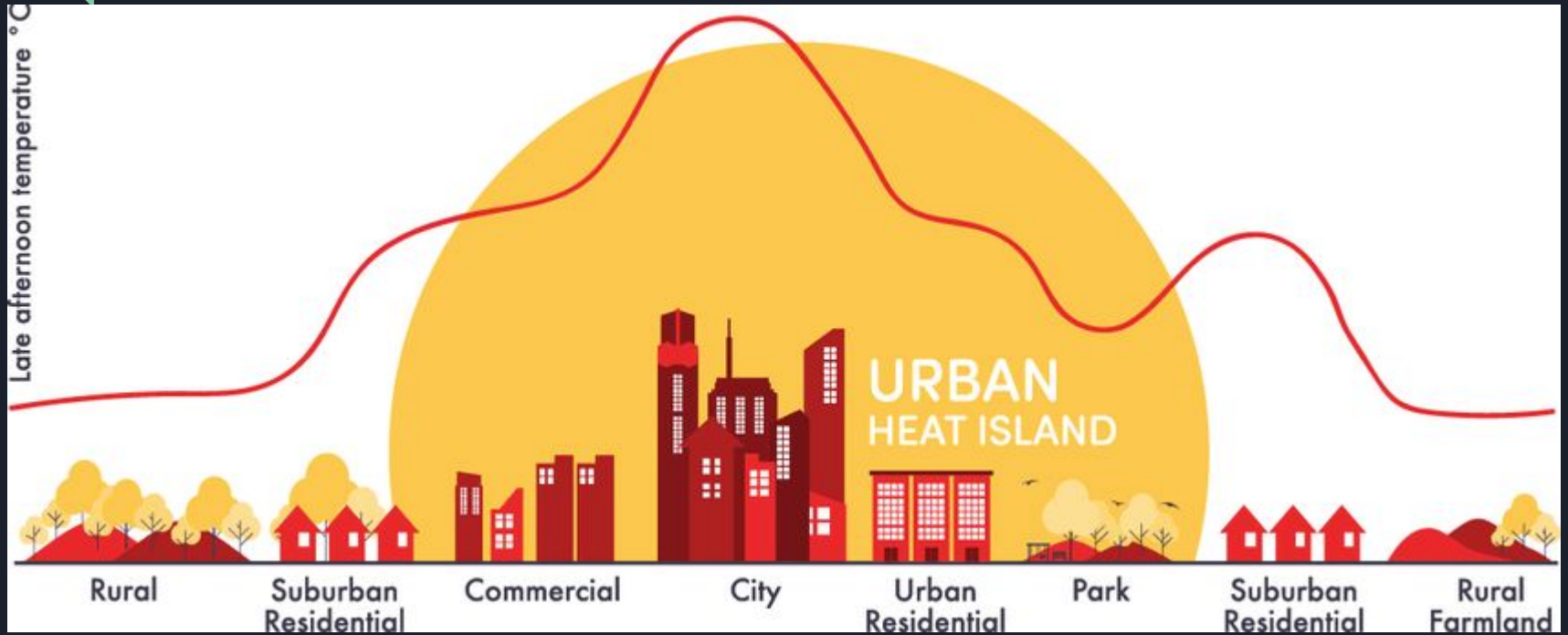
What? The EY Data Challenge has been operating since 2020 and is focused on the application of machine learning and statistical analysis to solve earth science problems

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 techniques into the creation of a regression model

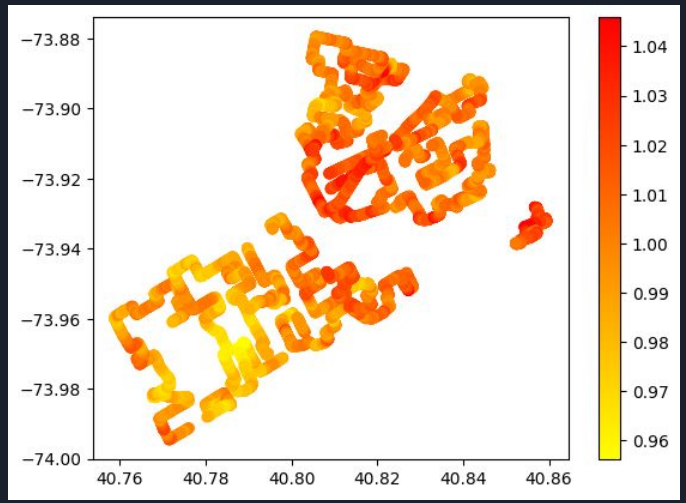
Urban Heat Islands Background Information



The Challenge

- Temperature data has been recorded at different locations across Manhattan and the Bronx in New York City on July 24th 2021 (training data on the right)
- The Urban Heat Index is essentially the temperature at one location compared to the average temperature of the city at a point in time
- Goal: Predict the UHI for a given location (longitude, latitude) from the test data which has locations without UHI values

	Longitude	Latitude	datetime	UHI Index
0	-73.919037	40.814292	24-07-2021 15:53	1.034616
1	-73.918978	40.814365	24-07-2021 15:53	1.028125
2	-73.918927	40.814433	24-07-2021 15:53	1.028125
3	-73.918875	40.814500	24-07-2021 15:53	1.025961
4	-73.918827	40.814560	24-07-2021 15:53	1.025961



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

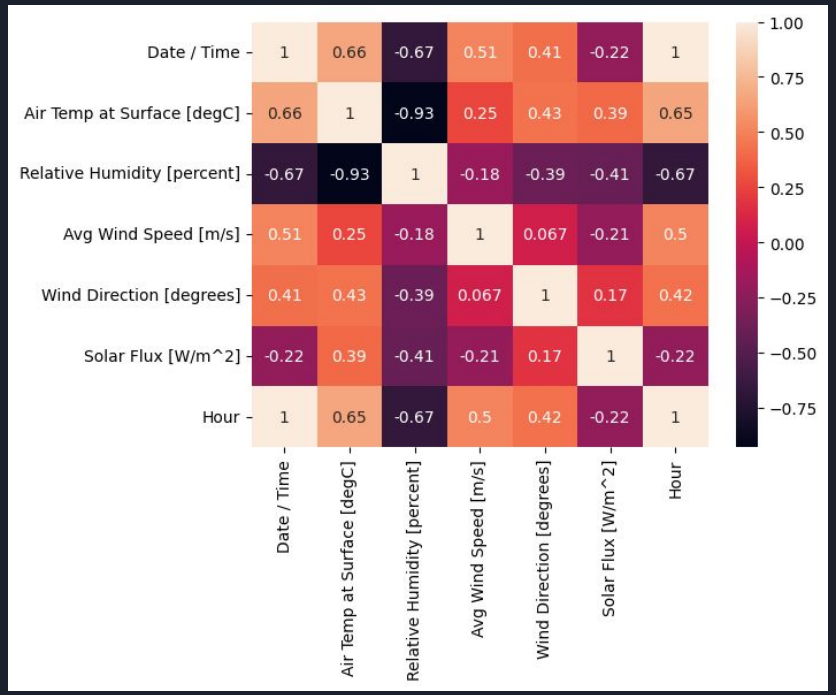
- Includes building height, roads, distance to shoreline, wind tunnels, tree canopy

EDA

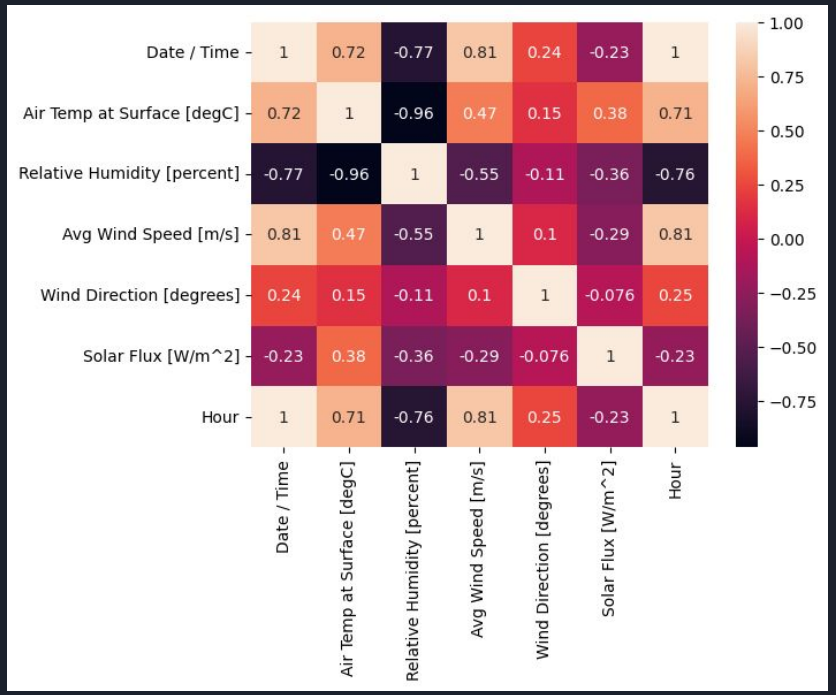
1. Looked at the correlation of the features.
2. Showcased Data Collected:
    a. UHI Index
    b. Air Temp at Surface [deg C]
    c. Relative Humidity [percent]
    d. Avg Wind Speed [m/s]
    e. Wind Direction [deg]
    f. Solar Flux [W/m^2]
3. Planetary Computer Data:
    a. Surface Temperature
 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

Correlation of Base Features

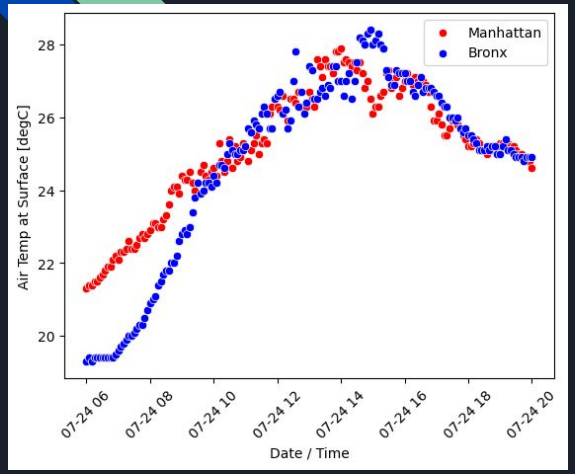
Manhattan Data

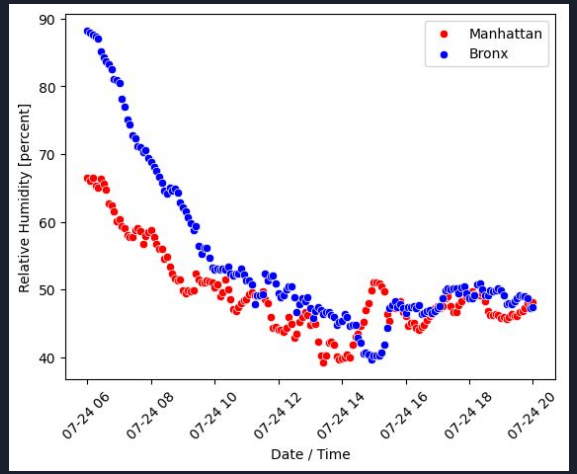


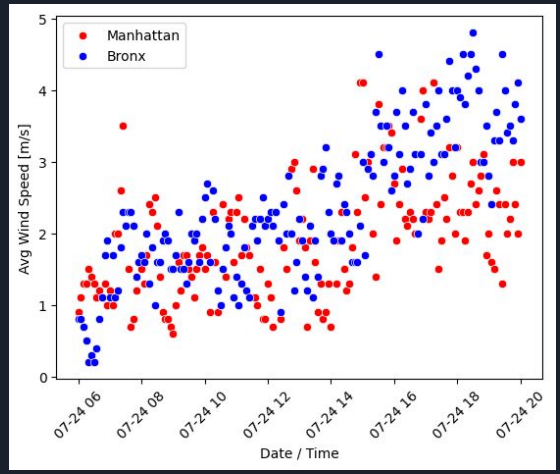
Bronx Data

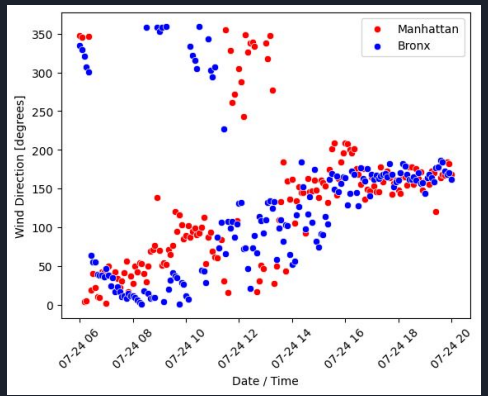


Manhattan Versus Bronx

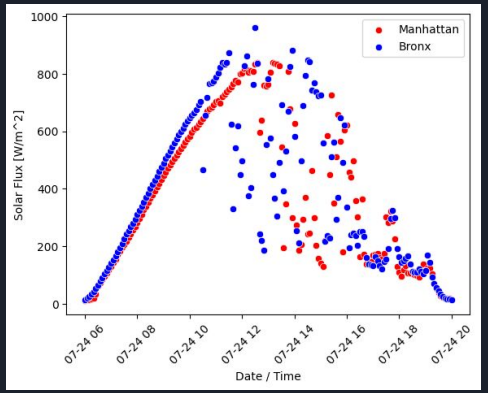








- Air Temp at Surface [deg]
- Relative Humidity [percent]
- Avg Wind Speed [m/s]
- Wind Direction [degrees]
- Solar Flux [W/m^2]



Planetary Computer

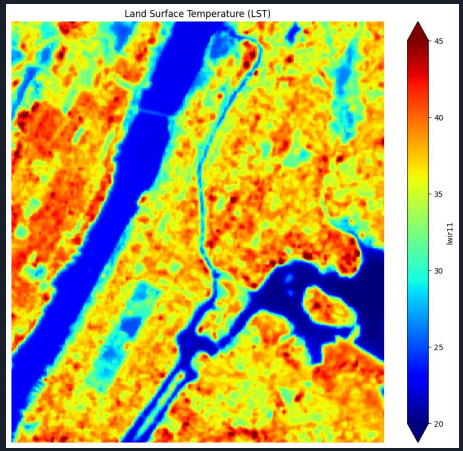
Microsoft's Planetary Computer is a large open source collection of different satellites and their remote sensing data of the Earth.

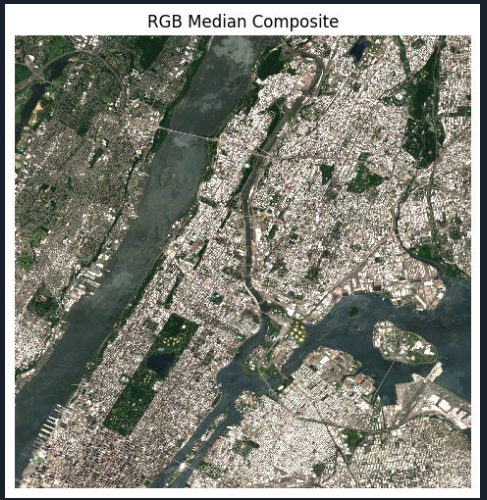
We will be mainly concentrating on

- Landsat (Land Surface Temperature) - 30/100 m per pixel
- Sentinel 2 - 10 m per pixel

Using the API we are able to construct time series of images to reduce cloud coverage or able to look at how temperature has changed using the landsat.

It is already clear that there are some locations in New York that have higher temperatures than others





How to construct different images

Depending on the satellite there are different bands (wavelengths).

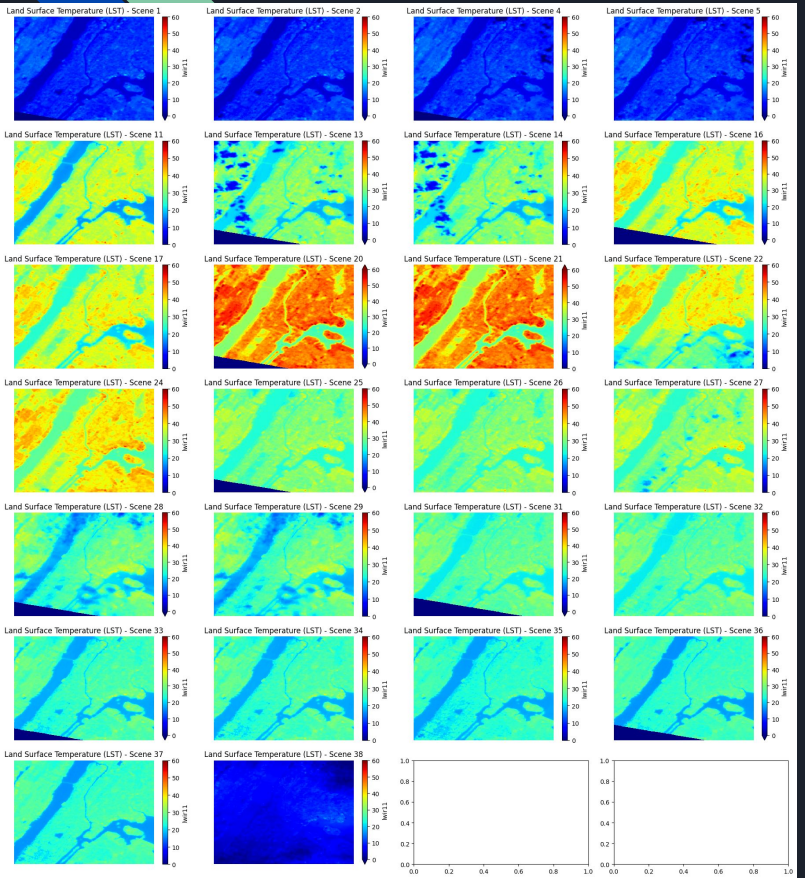
We are able to create data packages that we can call individual bands which are reflected visually.

This is going to be a key area of where we can feature engineer

One idea we have is looking at the greyscale of the bands or even performing a neural network for the ideal weights to predict temperature from Landsat

Sentinel-2 Bands	Central Wavelength (µm)	Resolution (m)
Band 1 - Coastal aerosol	0.443	60
Band 2 - Blue	0.490	10
Band 3 - Green	0.560	10
Band 4 - Red	0.665	10
Band 5 - Vegetation Red Edge	0.705	20
Band 6 - Vegetation Red Edge	0.740	20
Band 7 - Vegetation Red Edge	0.783	20
Band 8 - NIR	0.842	10
Band 8A - Vegetation Red Edge	0.865	20
Band 9 - Water vapour	0.945	60
Band 10 - SWIR - Cirrus	1.375	60
Band 11 - SWIR	1.610	20
Band 12 - SWIR	2.190	20

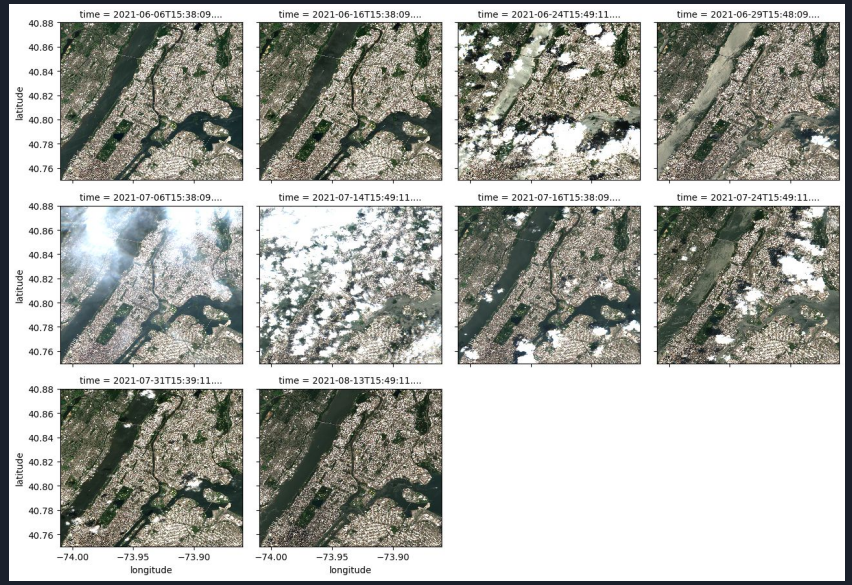
Planetary Computer Land Surface Temperature

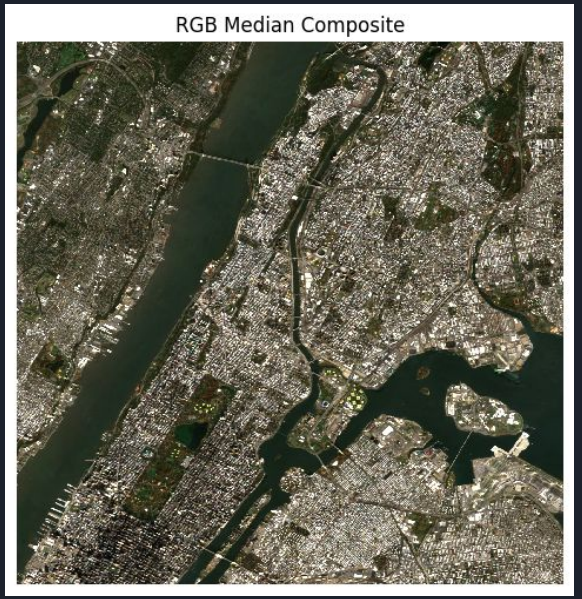


- Land surface temperature related to the detected surface temperature in 2024
- This was found using the Landsat Satellite

Planetary Median Composite

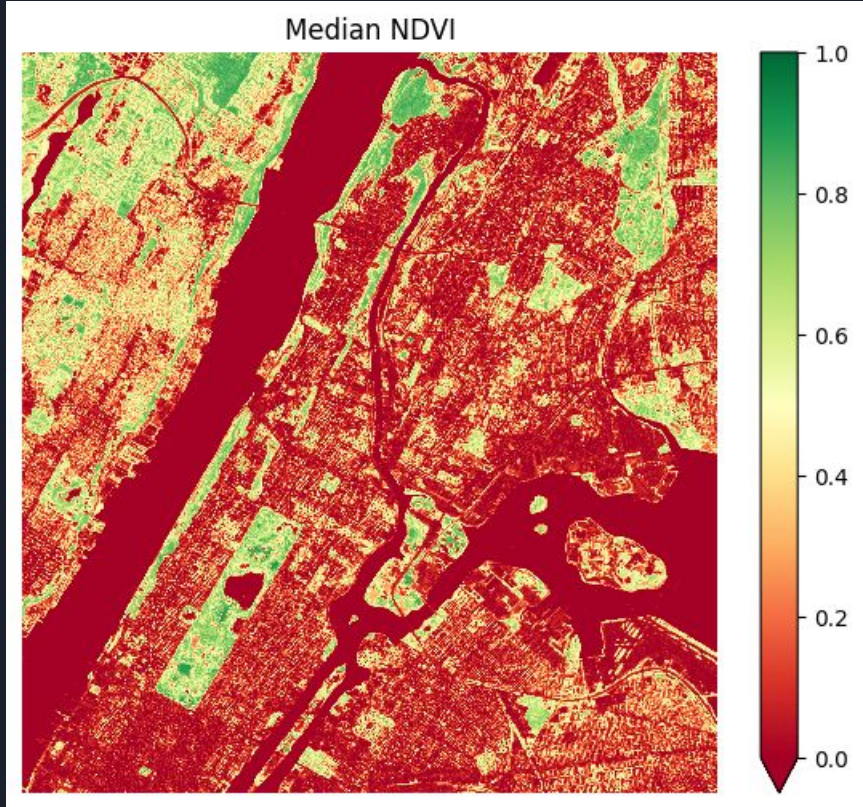
- Using normal xarray operations, we can compute the median pixel value at that point over many images.
- This chooses the middle value of the pixels which statistically removes cloud coverage





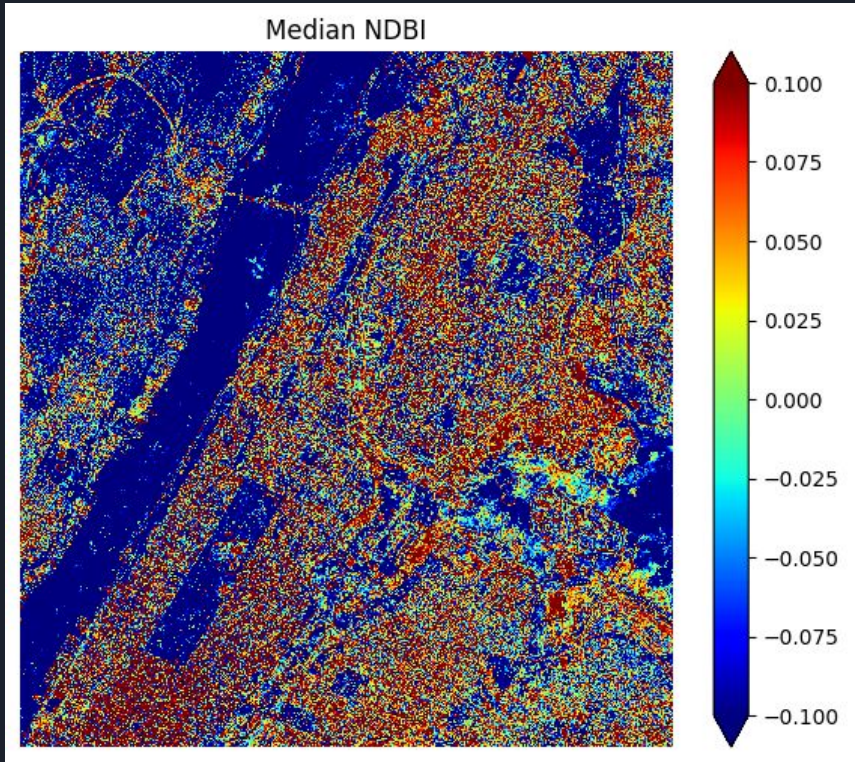
Planetary Computer NDVI

- 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.



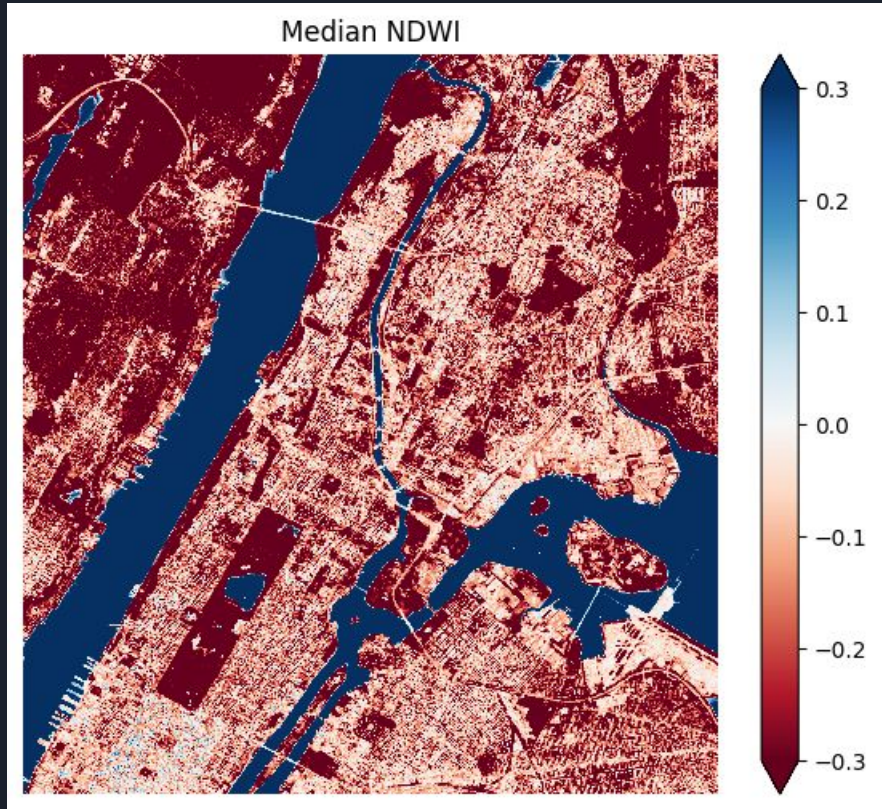
Planetary Computer NDBI

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



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.



Next steps

1. Explore different combination of remote sensing data to see which performs best
2. Explore how historical temperatures can be used to infer areas of concentration in building model
3. Incorporating data such as building density to understand how this can be used as a variable in our regression model

Any ideas???? Any questions??