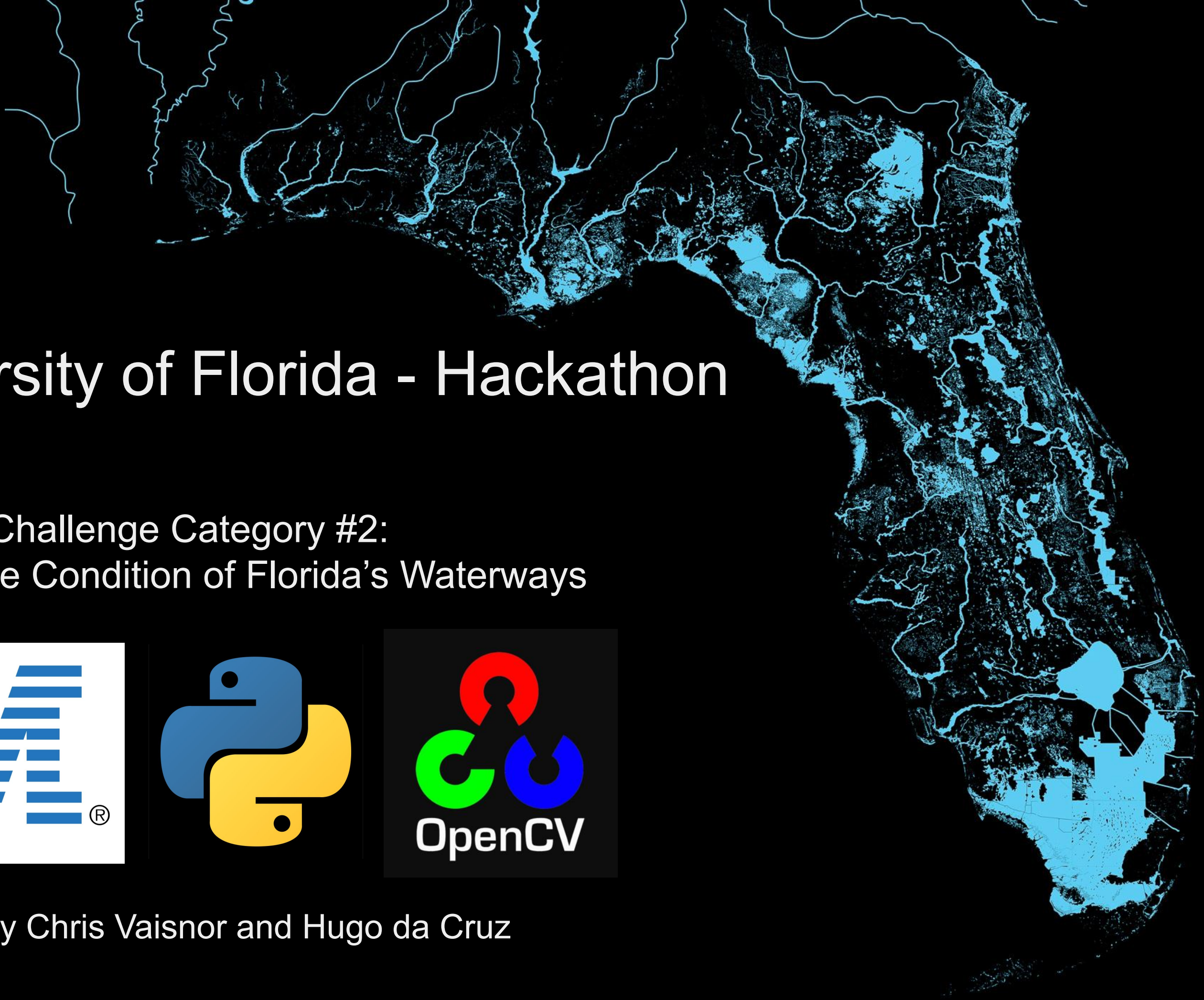
IBM & University of Florida - Hackathon

Challenge Category #2:
Improving the Condition of Florida's Waterways

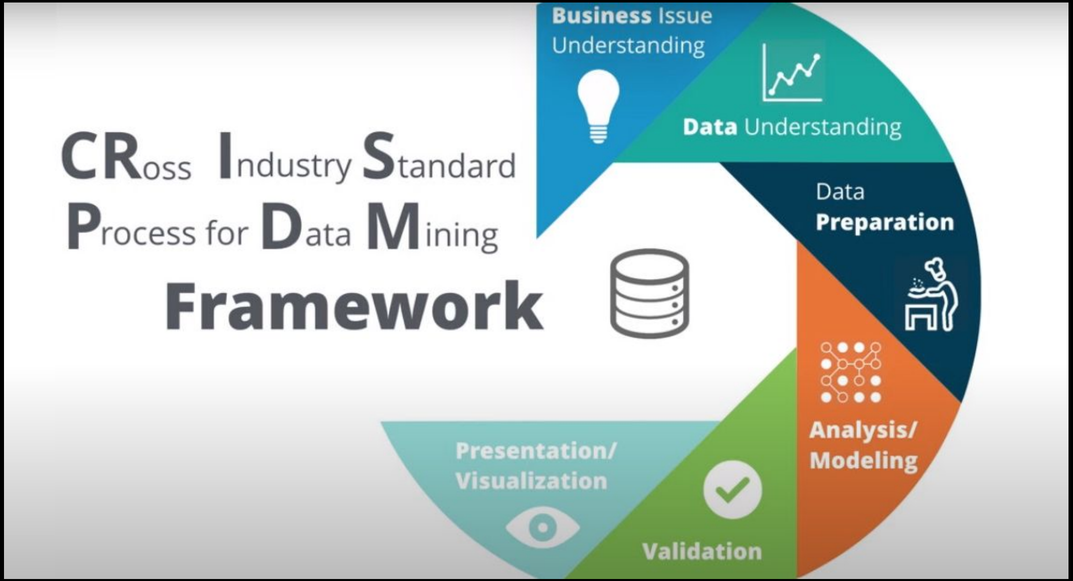


Project by Chris Vaisnor and Hugo da Cruz

Table of Contents

Our methodology is based on the CRISP-DM method for data science.

1. Understanding the Problem
2. Understanding the Data
3. Data Preparation
4. Applying Computer Vision
5. Evaluation
6. Deploying a Dashboard



Understanding the Problem, part 1:

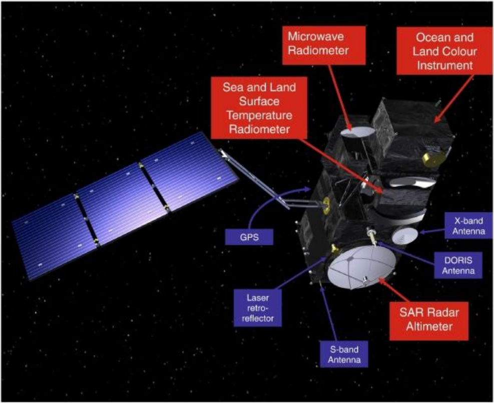
- What did we want to solve?
  - Florida has dangerous and costly algal blooms, and studying the raw data might give us insight that can help the communities along the waterways.
- What does success look like?
  - We would like to use algal bloom data to track their location and prevent costly health consequences. These consequences are for all species. Long term success would mean having higher quality data and over a longer period of time. In the short term, we would like a easy-to-use dashboard and notification system for the public. This system would warn locals about contaminants in the water and caution individuals of potential dangers.

(Manatees food sources are depleted by harmful algal blooms)

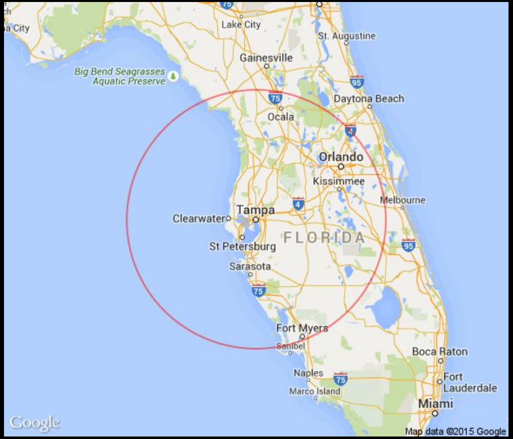
- What data is necessary?
  - Data showing location, size and intensity of harmful algal blooms is needed to perform meaningful analysis. We began searching for raw satellite images to detect cyanobacteria, one of the leading bacteria in harmful algal blooms. Other useful data would be the recordings of weather stations. This includes checking the ambient air temperature, humidity, UV index, and precipitation.

Understanding the Problem, part 2:

- What tools did we use?
  - The European Union launched a Sentinel-3 satellite mission with a specific Ocean and Land Color Instrument for taking photos of the Earth. This data is made publicly available and cataloged with the date it was observed. We planned to access these images and quantify the content.
- What was our final data resource?
  - We used prelabeled Sentinel-3 images offered by the National Oceanic and Atmospheric Administration (NOAA). The NOAA data viewer utilized offered image data specific to the western coast of Florida. We focused on the marine environment around the Tampa Bay area as it appears to often be affected by HABs.

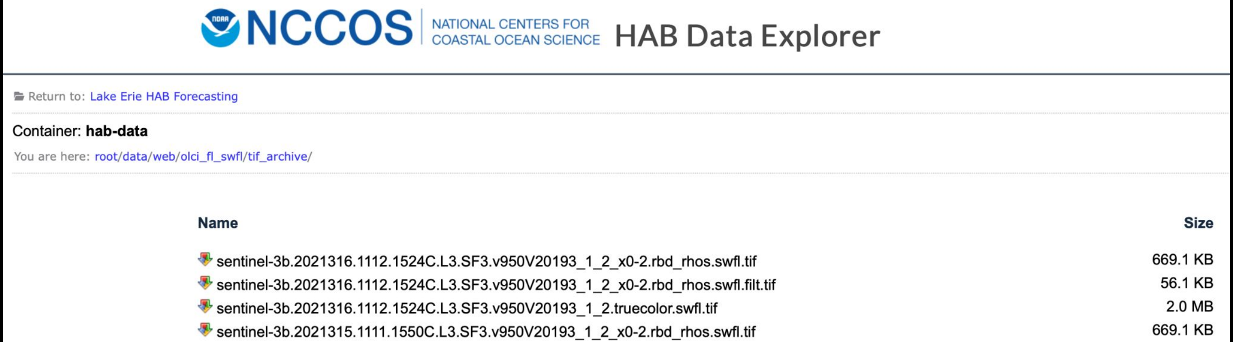


(Diagram of Sentinel-3)



Understanding the Data

- What did our dataset look like?
  - NOAA provided archived data from the Sentinel-3 mission that was labelled and organized by date. These were satellite images converted to .tif image files and cropped around the point of interest.
- Did the data have what we needed?
  - These image files contained a color spectrum corresponding to the concentration of chlorophyll fluorescence.
- What was the most important part of each observation?
  - We wanted to check the count, intensity, and location of the colors labeled as chlorophyll fluorescence.
- What data cleaning/transforming needs to happen?
  - The .tif file format is a loss-less image file, so there was not any noise in the data itself. We ran into an issue later that had to do with the quality of the actual data received by the satellite.



Return to: Lake Erie HAB Forecasting

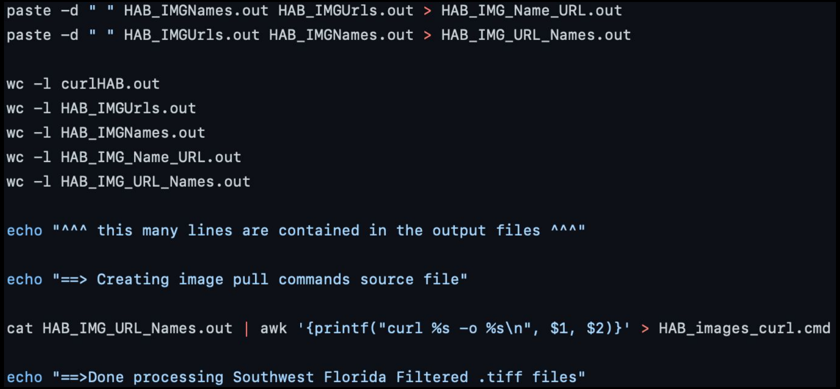
Container: **hab-data**

You are here: root/data/web/olci_fl_swfl/tif_archive/

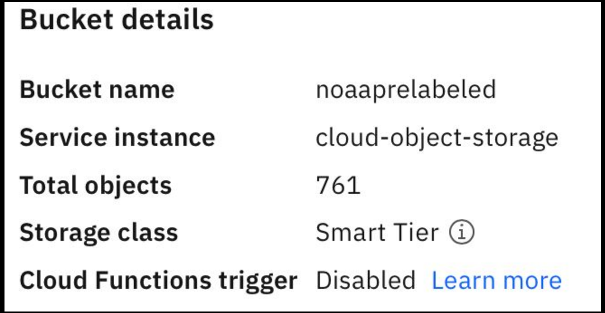
Name	Size
sentinel-3b.2021316.1112.1524C.L3.SF3.v950V20193_1_2_x0-2.rbd_rhos.swfl.tif	669.1 KB
sentinel-3b.2021316.1112.1524C.L3.SF3.v950V20193_1_2_x0-2.rbd_rhos.swfl.filt.tif	56.1 KB
sentinel-3b.2021316.1112.1524C.L3.SF3.v950V20193_1_2.truecolor.swfl.tif	2.0 MB
sentinel-3b.2021315.1111.1550C.L3.SF3.v950V20193_1_2_x0-2.rbd_rhos.swfl.tif	669.1 KB

Data Preparation

- What tools were used?
  - In order to pull the images from the NOAA website, we wrote a C shell script that had direct access to the image files and downloaded them to a local directory.
- Where is the data located and stored for accessibility?
  - Using the latest IBM Cloud technology, we utilized the straightforward API to upload the images directly to a Cloud Storage Object. A second C shell script was written to automate this process.
- What was the size of the data?
  - After web scraping and uploading to the IBM Cloud Object Storage, we had 761 image files with a total size of 50.9 MB. The historical image time period was between January 1, 2020 and mid October 2021.



```
paste -d " " HAB_IMGNames.out HAB_IMGUrls.out > HAB_IMG_Name_URL.out
paste -d " " HAB_IMGUrls.out HAB_IMGNames.out > HAB_IMG_URL_Names.out

wc -l curlHAB.out
wc -l HAB_IMGUrls.out
wc -l HAB_IMGNames.out
wc -l HAB_IMG_Name_URL.out
wc -l HAB_IMG_URL_Names.out

echo "^^^ this many lines are contained in the output files ^^^"

echo "==> Creating image pull commands source file"

cat HAB_IMG_URL_Names.out | awk '{printf("curl %s -o %s\n", $1, $2)}' > HAB_images_curl.cmd

echo "==>Done processing Southwest Florida Filtered .tiff files"
```

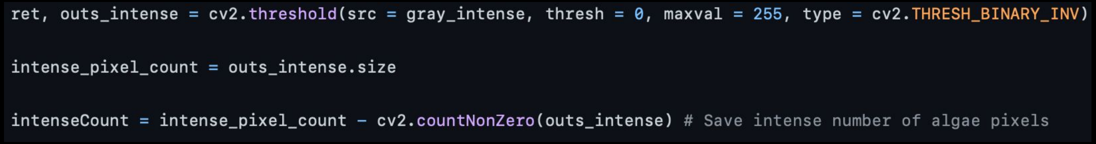


Bucket details

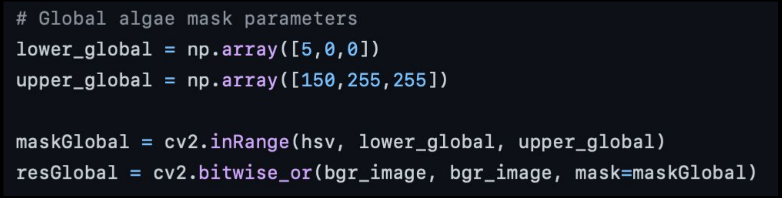
Bucket name	noaaprelabeled
Service instance	cloud-object-storage
Total objects	761
Storage class	Smart Tier ⓘ
Cloud Functions trigger	Disabled Learn more

Applying Computer Vision, part 1:

- What library did we use?
  - Given our familiarity with OpenCV, we decided to use the Python library to analyze images to detect algal blooms and their intensities.
- How did we transform the images for interpretation?
  - Using a Jupyter Notebook, we imported the image files with Blue-Green-Red color channels. In order to mask out specific pixels, we converted the image to HSV color channels to have better control over hue and saturation.
- How did we identify the harmful algal blooms?
  - After much trial and error, we found the correct parameters for masking out all the identified chlorophyll fluorescence, and then the correct parameters for the highest concentrations of the chlorophyll.



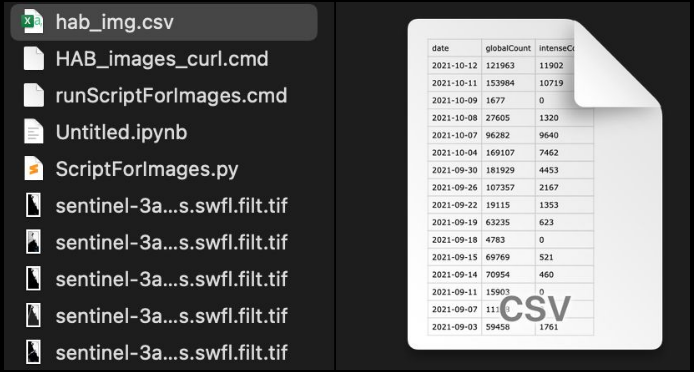
```
ret, outs_intense = cv2.threshold(src = gray_intense, thresh = 0, maxval = 255, type = cv2.THRESH_BINARY_INV)

intense_pixel_count = outs_intense.size

intenseCount = intense_pixel_count - cv2.countNonZero(outs_intense) # Save intense number of algae pixels
```



```
# Global algae mask parameters
lower_global = np.array([5,0,0])
upper_global = np.array([150,255,255])

maskGlobal = cv2.inRange(hsv, lower_global, upper_global)
resGlobal = cv2.bitwise_or(bgr_image, bgr_image, mask=maskGlobal)
```

Applying Computer Vision, part 2:

- What did we want to track about each image?
  - For each image data observation, we counted the pixels with any detection of chlorophyll, and also counted the pixels that fall under the highest levels of concentration.
- How did we systematically sort through each image and quantify the values?
  - After testing our masking techniques in a Jupyter Notebook, we converted it to a Python script. Then we wrote a C shell script that wrapped over the Python script and iterated through the directory of image files.
  - The quantified dataset was a .csv file that we appended the variables of each image to. This would allow us to use the Pandas Library for data analysis.



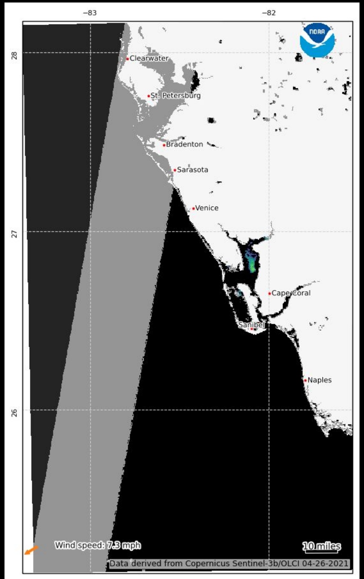
hab_img.csv
HAB_images_curl.cmd
runScriptForImages.cmd
Untitled.ipynb
ScriptForImages.py
sentinel-3a...s.swfl.filt.tif
sentinel-3a...s.swfl.filt.tif
sentinel-3a...s.swfl.filt.tif
sentinel-3a...s.swfl.filt.tif
sentinel-3a...s.swfl.filt.tif

date	globalCount	intenseC
2021-10-12	121963	11902
2021-10-11	153984	10719
2021-10-09	1677	0
2021-10-08	27605	1320
2021-10-07	96282	9640
2021-10-04	169107	7462
2021-09-30	181929	4453
2021-09-26	107357	2167
2021-09-22	19115	1353
2021-09-19	63235	623
2021-09-18	4783	0
2021-09-15	69769	521
2021-09-14	70954	460
2021-09-11	15903	0
2021-09-07	[illegible]	[illegible]
2021-09-03	59458	1761

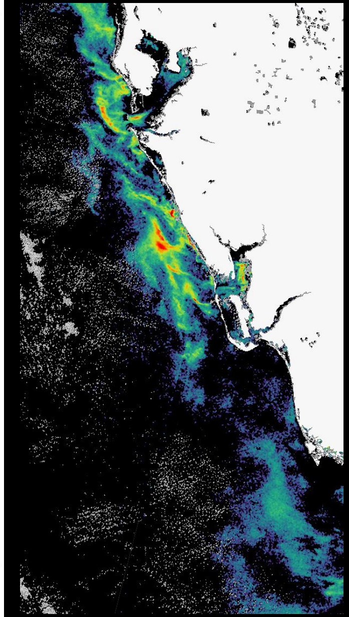
Evaluation, part 1:

- Results vs. Expectations?
  - We expected results with small variations in algal bloom signatures and discovered that cloud cover off the coast of Tampa Bay was rather prevalent and obscured algal bloom signatures. We also encountered satellite images that were not complete. We presume that the data was not completely sent back to Earth stations or that the orbit of the satellite does not allow for consistent image crops.
- Is there a benchmark to compare images?
  - We have high quality images that are not affected by cloud quality that are also absent of algal bloom signatures as benchmarks. Images with algal bloom signatures were compared against benchmark images to help isolate signatures and quantify intensity of algal blooms.

(Left image is an example of a bad observation)

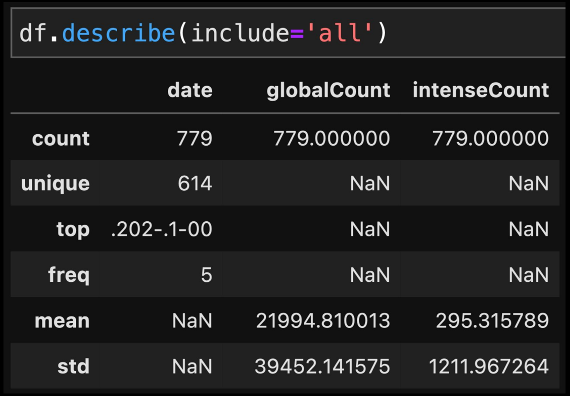


Red Band Difference (RBD) showing relative fluorescence from high (red) to low (violet). A median filter was applied to remove speckle. Winds from NOAA NDBC station VENF1.



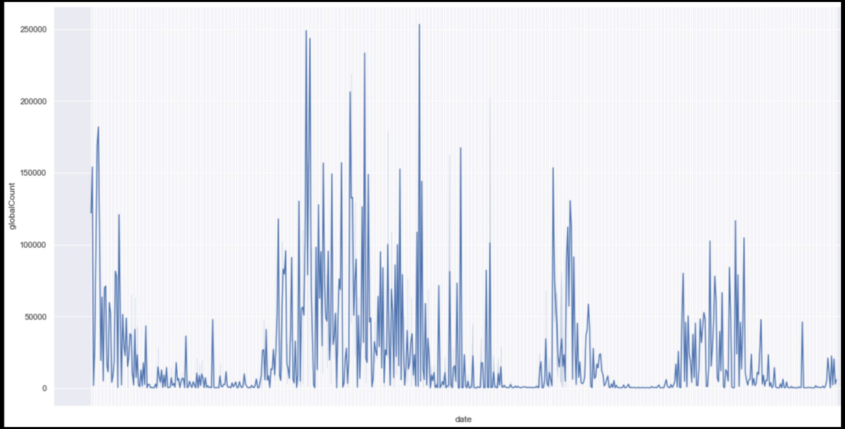
Evaluation, part 2:

- Summary statistics?
    - Our summary statistics show a very high standard deviation. The SD is almost twice the mean for the globalCount variable. This shows that there is a higher level of algae count variability compared to the average count across all the images.
- Can our results be improved?
    - Our results could have been more definitive in the absence of cloud cover and missing satellite image data. Having more images available over a longer period of time would have allowed us to show seasonal as well as decadal variations.

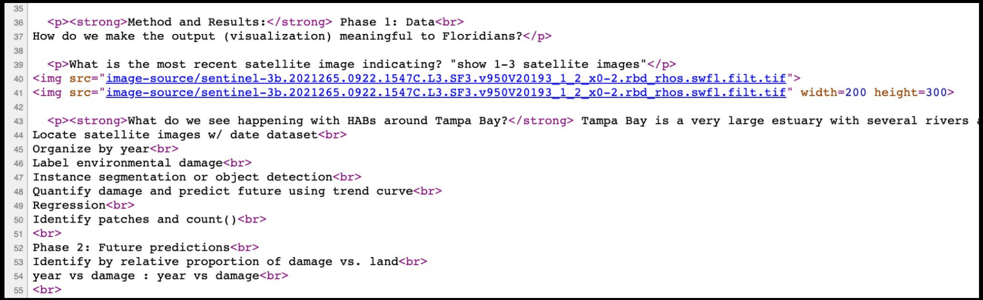


```
df.describe(include='all')
```

	date	globalCount	intenseCount
count	779	779.000000	779.000000
unique	614	NaN	NaN
top	.202-.1-00	NaN	NaN
freq	5	NaN	NaN
mean	NaN	21994.810013	295.315789
std	NaN	39452.141575	1211.967264



Deploying a Dashboard



```html

  <p><strong>Method and Results:</strong> Phase 1: Data<br>
How do we make the output (visualization) meaningful to Floridians?</p>

  <p>What is the most recent satellite image indicating? "show 1-3 satellite images"</p>
<img src="image-source/sentinel-3b.2021265.0922.1547C.L3.SF3.v950V20193_1_2_x0-2.rbd_rhos.swfl.filt.tif">
<img src="image-source/sentinel-3b.2021265.0922.1547C.L3.SF3.v950V20193_1_2_x0-2.rbd_rhos.swfl.filt.tif" width=200 height=300>

  <p><strong>What do we see happening with HABs around Tampa Bay?</strong> Tampa Bay is a very large estuary with several rivers
Locate satellite images w/ date dataset<br>
Organize by year<br>
Label environmental damage<br>
Instance segmentation or object detection<br>
Quantify damage and predict future using trend curve<br>
Regression<br>
Identify patches and count()<br>
<br>
Phase 2: Future predictions<br>
Identify by relative proportion of damage vs. land<br>
year vs damage : year vs damage<br>
<br>
```

- What resources did we consider?
  - There are a number of aesthetically pleasing exploratory data analysis applications that we could use to present our results. We chose a more traditional HTML page framework, however, we had grand aspirations to implement an interactive dashboard.
- Presenting our findings:
  - We built a website to show the three latest satellite images and provide actionable information to the public. The website includes background information about harmful algal blooms and why we need to be paying attention to them.
- What would we include if we had more funding/time?
  - We would like to expand our exploratory data analysis visualizations on this dashboard. That includes using color distinctions with multiple plotting types. We want to make our dashboard a compelling resource for the local communities and present our information in the most professional way possible.