HW7

March 27, 2022

0.1 HW 7: Data Description & Preprocessing with Input Data Visualization

0.1.1 OCEN 460

0.1.2 Team: _/Sample_Text/

0.1.3 Members: Nate Baker and James Frizzell

[1]:

```
import pandas as pd
import matplotlib.pyplot as plt
import pathlib
import os

%matplotlib inline
#Github: https://github.com/jafrizzell/coral-prediction.git
#Describe Datasets and project idea
```

- The World Ocean Atlas (WOA) data cannot be shown in it's raw for here because it is too large to be shared in teh github repository where the data is stored. The metadata for it looks as follows. Replace temperature with "salinity" or "dissolved oxygen" for the other two datasets collected from WAO.

Latitude | Longitude | Temperature@0m depth (Celsius) | Temp@5m | Temp@10m | Temp@15m |…| Temp@5500m

- The Deep Sea Coral Data (DSC) has the following metadata

Latitude | Longitude | Depth (m) |

0.1.4 1. The deep sea coral dataset reports latitude and longitude of known coral growth locations with the depth at which the coral is growing. The World Ocean Atlas reports depth measurements in increments of 5 meters for depths of 0 to 100 meters, 10 meters for 100 to 500 meters, 50 meters for 500 to 2000 meters, and in 100 meters for greater than 2000 meters. The following code was used and adjusted to round the Deep Sea Coral dataset to match this convention.

path = 'C:/Users/jafri/Documents/GitHub/coral-prediction/processed_data/deep_sea_corals_rounded.csv'

raw = pd.read_csv(path)

def round_depth(x, base): return int(base * round(float(x)/base))

raw['depth'] = raw['depth'].apply(lambda x: round_depth(x, base=5)) raw = raw[raw.depth >= 0] print(len(raw)) raw.to_csv(‘C:/Users/jafri/Documents/GitHub/coral-prediction/processed_data/deep_sea_corals_rounded_depthcorr.csv’)

0.1.5 2. The following code aligns latitude and longitude values from the Deep Sea Coral dataset with the lat/long values from the WOA dataset with a tolerance of 0.5 degrees. Second_param file can be changed to indicate the oceanographic variable of interest. WOA data is right-joined to DSC data for further preprocessing.

0.1.6 The code yields a .csv file that contains the DSC data and the WOA data. The WOA data is depth-stratified.

import geopandas

coral = ‘D:/TAMU Work/TAMU 2022 SPRING/OCEN 460/depthtempsal_short2.csv’ second_param = ‘D:/TAMU Work/TAMU 2022 SPRING/OCEN 460/woa18_all_O00mn01.csv’ # “O00mn01” indicates O2 data

raw_coral = pd.read_csv(coral) raw_coral = geopandas.GeoDataFrame(raw_coral, geometry=geopandas.points_from_xy(raw_coral.longitude, raw_coral.latitude)) raw_coral.depth = raw_coral.depth.astype(float) raw_coral.latitude = raw_coral.latitude.astype(float) raw_coral.longitude = raw_coral.longitude.astype(float)

raw_param = pd.read_csv(second_param) raw_param = raw_param.astype(float) raw_param = geopandas.GeoDataFrame(raw_param, geometry=geopandas.points_from_xy(raw_param.longitude, raw_param.latitude))

depth_sal = raw_coral.sjoin_nearest(raw_param, max_distance=0.5)

depth_sal.to_csv(‘D:/TAMU Work/TAMU 2022 SPRING/OCEN 460/depthtempsaloxy.csv’, index=False)

0.1.7 3. To resolve the stratified nature of the WOA data, the following code is used to select the corresponding WOA column for the DSC depth of interest.

path = ‘D:/TAMU Work/TAMU 2022 SPRING/OCEN 460/depthtempsaloxy.csv’

raw = pd.read_csv(path) raw = raw[raw[‘depth’] <= 5500] skipped = 0 for i in range(len(raw)): try: depth = str(raw[‘depth’][i]) raw[‘oxygen’][i] = raw[depth][i] except KeyError: skipped+=1 pass

print(“skipped:”, skipped)

raw.to_csv(‘D:/TAMU Work/TAMU 2022 SPRING/OCEN 460/depthtempsaloxy_short.csv’, index=False)

0.1.8 4. The following code determines the maximum depth for each lat/long pair in the WOA dataset. These datapoints were then used to create a control dataset describing where coral is not present, in order to compare to the DSC dataset. Code displayed in sections 2 and 3 were used to add the temperature, salinity, and oxygen variables to the control dataset.

path = ‘D:/TAMU Work/TAMU 2022 SPRING/OCEN 460/woa18_decav_t00mn04.csv’

raw = pd.read_csv(path) depth = []

for i in range(len(raw)): for j in range(103): curr = raw.iloc[i, -1-j] plus = raw.iloc[i, -2-j if j == 0 and np.isfinite(curr): depth.append(raw.columns[-1]) break elif np.isnan(curr) and np.isfinite(plus): depth.append(raw.columns[-2-j]) break elif j == 102: depth.append(raw.columns[-1]) print(len(depth)) print(len(raw.latitude)) out = pd.DataFrame({'latitude': raw.latitude, 'longitude': raw.longitude, 'depth': depth})

out.to_csv('D:/TAMU Work/TAMU 2022 SPRING/OCEN 460/depths.csv', index=False)

Ultimately, the final dataset had the following metadata

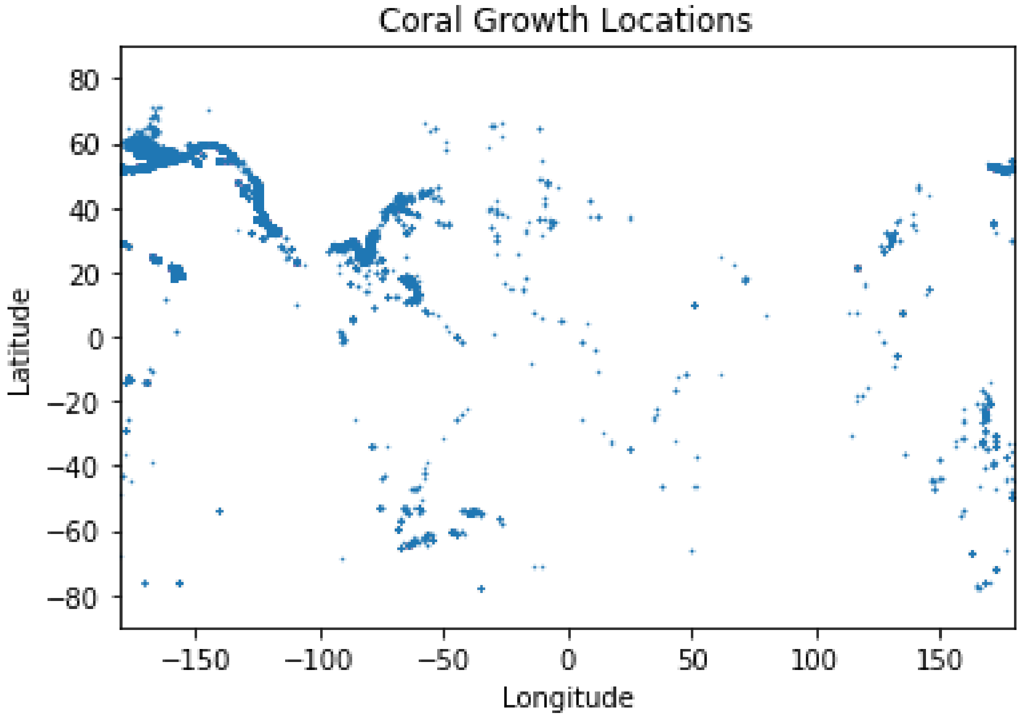
Latitude | Longitude | Depth (m) | Temperature (c) | Salinity (ppt) | Dissolved Oxygen (umol/kg)

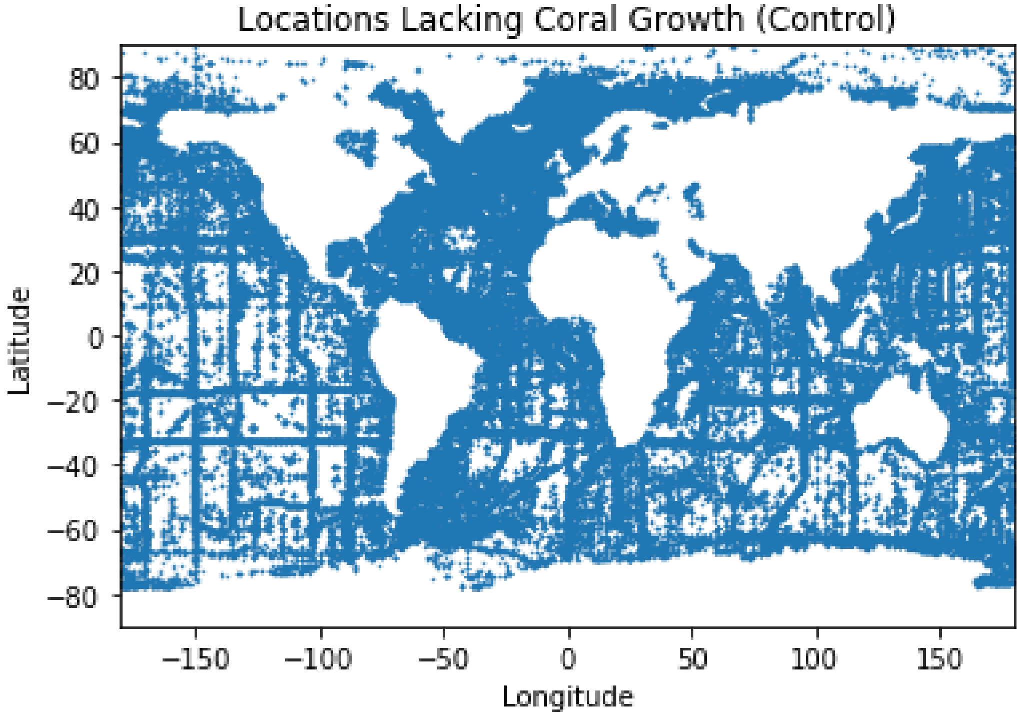
0.1.9 5. The following code visualizes the two datasets.

[2]:

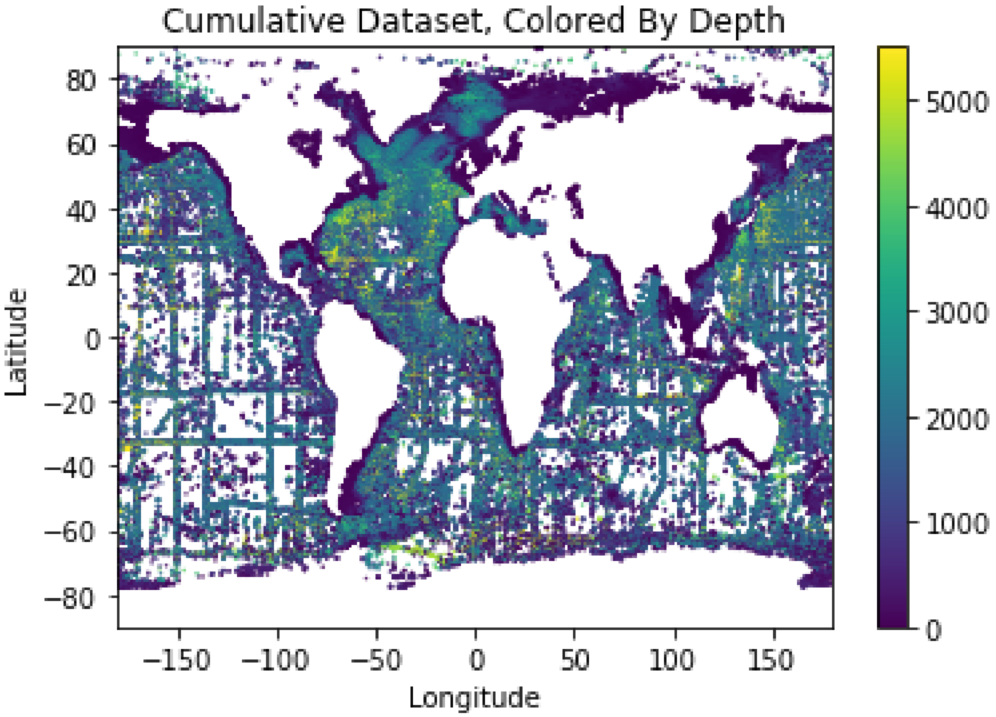
```
#Reprocessed Data for Visualization
path = str(pathlib.Path(os.getcwd())) +␣
 ↪'\processed_data\combined_data_truncated.csv'
raw = pd.read_csv(path)
print(raw.describe())

#Visualization
coral_present_bool = raw[raw.coral_present == 1]
plt.scatter(coral_present_bool['longitude'], coral_present_bool["latitude"], s=␣
 ↪0.2)
plt.title("Coral Growth Locations")
plt.xlabel("Longitude")
plt.ylabel("Latitude")
plt.xlim([-180,180])
plt.ylim([-90,90])
plt.show()

coral_missing_bool = raw[raw.coral_present == 0]
plt.scatter(coral_missing_bool['longitude'], coral_missing_bool["latitude"], s=␣
 ↪0.2)
plt.title("Locations Lacking Coral Growth (Control)")
plt.xlabel("Longitude")
plt.ylabel("Latitude")
plt.xlim([-180,180])
plt.ylim([-90,90])
plt.show()

print("Number of Coral Growth Datapoints:", len(coral_present_bool))
print("Number of Datapoints with no Coral Growth", len(coral_missing_bool))

plt.scatter(raw.longitude, raw.latitude, s=0.2, c=raw.depth)
plt.title("Cumulative Dataset, Colored By Depth")
plt.xlabel("Longitude")
plt.ylabel("Latitude")
plt.xlim([-180,180])
plt.ylim([-90,90])
plt.colorbar()
plt.show()
plt.scatter(raw.longitude, raw.latitude, s=0.2, c=raw.temperature)
plt.title("Cumulative Dataset, Colored By Temperature")
plt.xlabel("Longitude")
plt.ylabel("Latitude")
plt.xlim([-180,180])
plt.ylim([-90,90])
plt.colorbar()
plt.show()
plt.scatter(raw.longitude, raw.latitude, s=0.2, c=raw.salinity)
plt.title("Cumulative Dataset, Colored By Salinity")
plt.xlabel("Longitude")
plt.ylabel("Latitude")
plt.xlim([-180,180])
plt.ylim([-90,90])
plt.colorbar()
plt.show()
plt.scatter(raw.longitude, raw.latitude, s=0.2, c=raw.oxygen)
plt.title("Cumulative Dataset, Colored By Oxygen")
plt.xlabel("Longitude")
plt.ylabel("Latitude")
plt.xlim([-180,180])
plt.ylim([-90,90])
plt.colorbar()
plt.show()
```

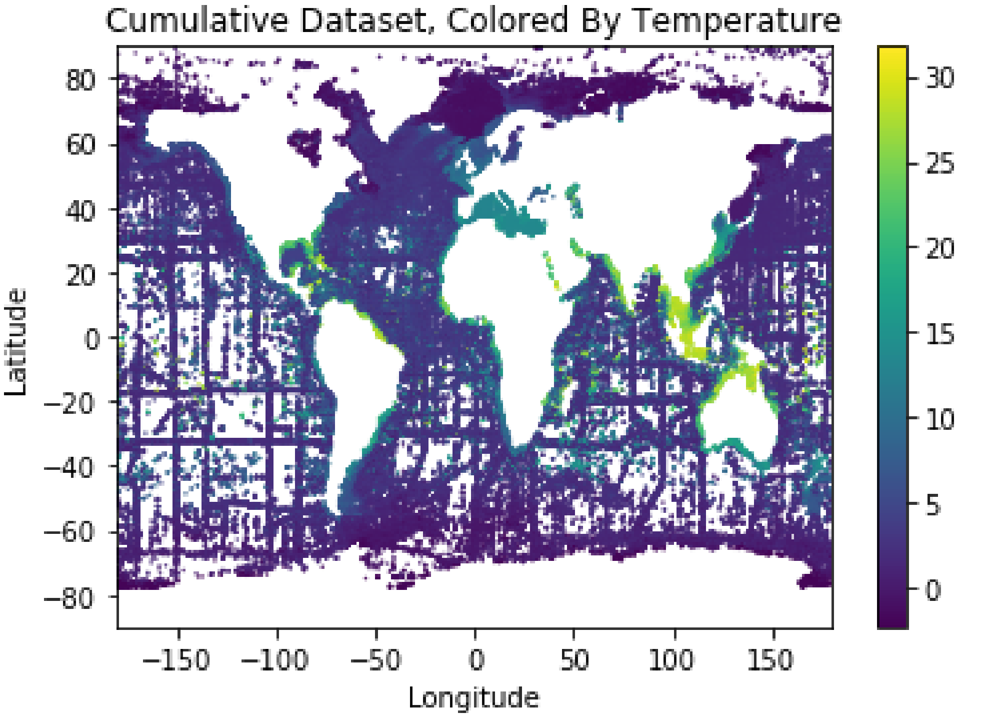
```
       coral_present       latitude      longitude          depth  \
count  341773.000000  341773.000000  341773.000000  341773.000000
mean        0.582340      23.237462     -68.404081    1066.013070
std         0.493174      33.783962      96.027313     989.540361
min         0.000000     -77.875000    -179.989750       0.000000
25%         0.000000      16.625000    -124.339620     235.000000
50%         1.000000      35.641580    -119.498760     850.000000
75%         1.000000      40.811190     -25.625000    1750.000000
max         1.000000      89.875000     179.989750    5500.000000

             round_d    temperature       salinity         oxygen
count  341773.000000  341773.000000  341773.000000  341773.000000
mean     1067.012915       5.191193      34.392171      42.294298
std       989.771161       4.584556       1.343642      29.076344
min         0.000000      -2.271000       0.000000       0.199000
25%       225.000000       2.452000      34.228000      12.890000
50%       850.000000       3.878000      34.520000      41.558000
75%      1750.000000       7.371000      34.699000      65.495000
max      5500.000000      31.751000      41.310000     132.182000
```

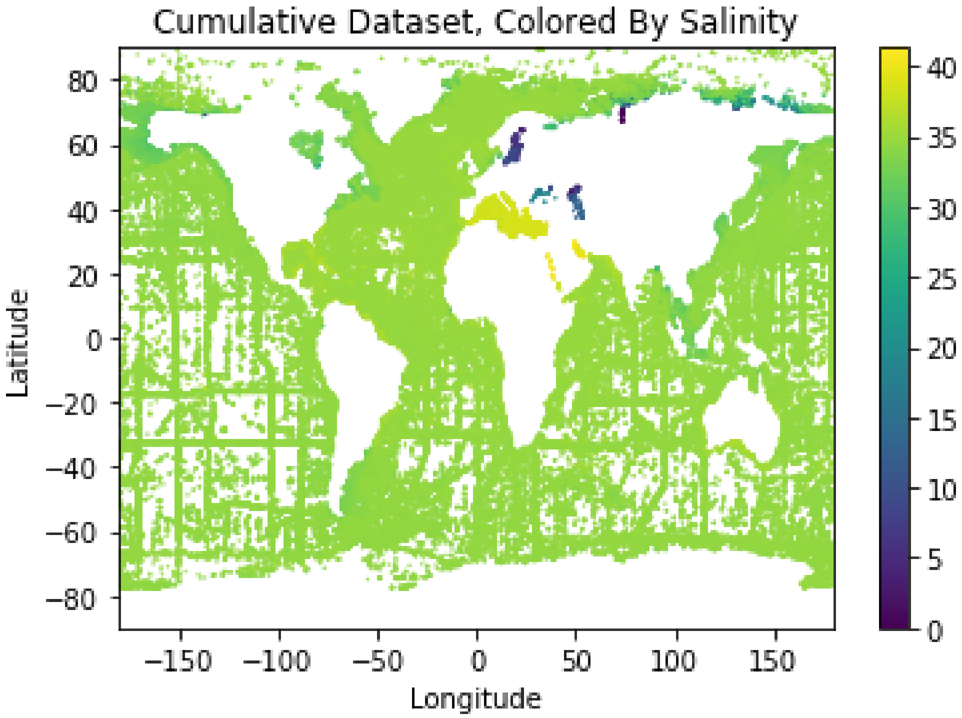

Coral Growth Locations
80
60
40
20
0
−20
−40
−60
−80
Latitude
−150
−100
−50
0
50
100
150
Longitude

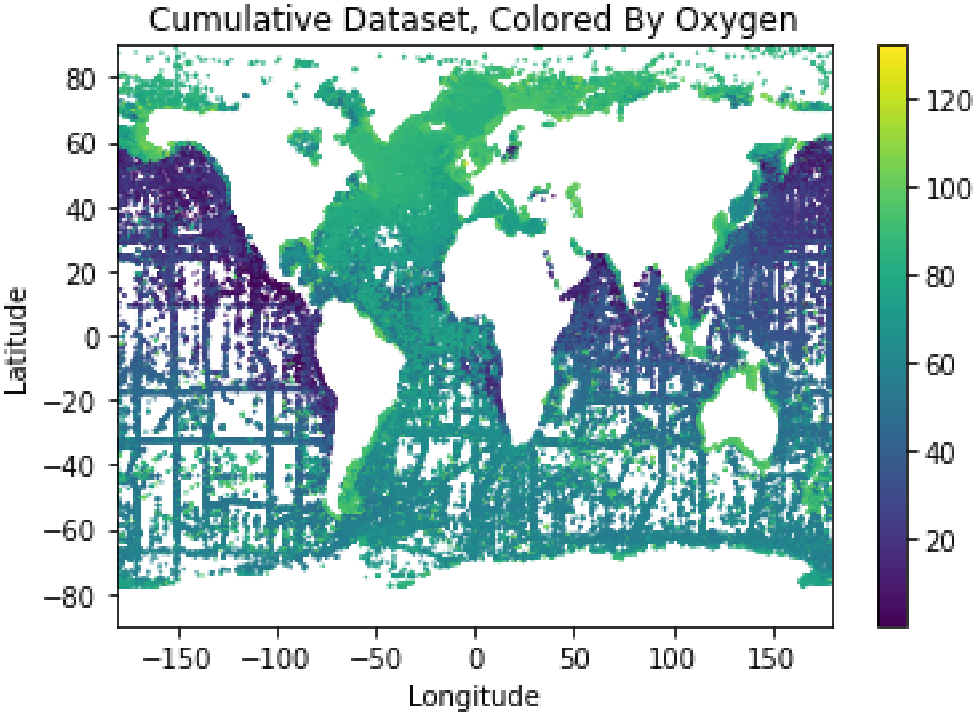

Locations Lacking Coral Growth (Control)
80
60
40
20
0
−20
−40
−60
−80
Latitude
−150
−100
−50
0
50
100
150
Longitude

```
Number of Coral Growth Datapoints: 199028
Number of Datapoints with no Coral Growth 142745
```






Cumulative Dataset, Colored By Salinity
80
60
40
20
0
−20
−40
−60
−80
Latitude
−150
−100
−50
0
50
100
150
Longitude
40
35
30
25
20
15
10
5
0


Cumulative Dataset, Colored By Oxygen
80
60
40
20
0
−20
−40
−60
−80
Latitude
−150
−100
−50
0
50
100
150
Longitude
120
100
80
60
40
20

[ ]:

```
```