Milestone 8: Final project

Project Introduction:

Our group's goal is to illustrate the domino effect that overfishing can have on the ecosystem. We hypothesize that as fishing activity increases, harmful algae blooms become more prevalent, leading to the phenomenon of coral reef "bleaching" or partial death. We chose this issue because we all have a deep connection with our environment and are committed to preserving it. To test our hypothesis regarding the correlation between commercial fishing capture and harmful marine environmental trends, we analyzed data related to Karenia brevis algae cell count, aggregate commercial fishing capture, and coral reef bleaching.

By examining the relationship between overfishing, harmful algae, and coral reef health, we hope to shed light on the broader impacts of human activity on our planet's delicate balance.

Data cleaning:

Ethan:

- Harmful Karenia brevis algae cell count
- source: https://www.ncei.noaa.gov/maps/habsos/maps.htm
- CVS file - original size (190339, 25)
- Multitude of columns including: STATE_ID, DESCRIPTION, LATITUDE, LONGITUDE, SAMPLE_DATE, SAMPLE_DEPTH, GENUS, SPECIES, CATEGORY, CELLCOUNT
- Focused on a few columns dropping WIND_DIR', 'WIND_DIR_UNIT', 'WIND_DIR_QA', 'WIND_SPEED', 'WIND_SPEED_UNIT', 'WIND_SPEED_QA'
- Cleaned date information, removing outliers by determining invalid dates
- Took awhile to find physical data to use, most related topics were research papers with no data publicly available

Khizer:

- Sql file original size: (34846,11)
- Had to write extensive queries to extract the data related to the project
- Extensive cleaning was involved mostly through sql queries and some using pandas dataframe
- Ocean_Name, Country_Name, Date_Year, Data_Source, Bleaching_Prevalence_Score, ClimSST, Windspeed, SSTA, SSTA_DHW, TSA, TSA_DHW are the values that were extracted through queries

- Tables used: Ocean, Country, Site_Info, Bleaching_tbl, Environmental_tbl
- I combined data from ethan and sean to find some correlations between my data and theirs (Had to normalize the data to a scale of 0-3). Ended up finding some correlations but more research is required.

Sean:

- CSV file original size: (53,6)
- I used 5 separate csv files of the same size to gather all Gulf Region data
- This files are small due to containing annual data
- Minimal cleaning was required. Just dropped irrelevant data columns and formatted numeric and date types
- Values consisted of date, region name, pounds, dollars collection type, and metric tons
- I also combined my data with a dataset (24025,3) that Ethan cleaned. In an effort to perform linear regression on this data, I formatted the date column and found the mean cell count for each year. Essentially turning the dataset into annualized cell count data. This allowed me to combine it with my annual fishing capture data and analyze the relationship between cell count and aggregate fish capture.

Exploratory data analysis:

Descriptions and visualizations highlighting our exploratory data analysis can be found in the Google Collab folders located in our GitHub repository.

Data:

Data files can be found in the Google Collab folders located in our GitHub repository.

ML/Stats:

Machine learning and their associated statistics can be found in the Google Collab folders located in our GitHub repository.

Results:

Despite our efforts, we did not observe a strong direct correlation between

commercial overfishing and the specific harmful environmental impacts that we were analyzing. Our findings suggest that there are likely multiple factors contributing to these negative trends, including those beyond the scope of our research, such as climate change and human-related pollution from sources like oil spills and agricultural runoff. Gathering relevant data posed a challenge, as the data on commercial fish capture fluctuated significantly from year to year and aligning our data to a specific region of the Gulf coast proved difficult. Additionally, working with data of varying time intervals (annually, weekly, monthly) made it challenging to model and identify trends.

In retrospect, analyzing the impacts of climate change and pollution may have yielded more insights into the causes of the negative marine environmental trends we observed. Unfortunately, publicly available data on agricultural runoff and oil spill pollution was limited. Nonetheless, our research provided valuable insight into the complexity of the issue and identified potential paths for further exploration.

In summary, while we did not identify a singular factor driving the environmental changes in the Gulf region, our research helped us gain a better understanding of the issue and highlighted the need for continued exploration of potential contributing factors.

References:

https://www.ncei.noaa.gov/maps/habsos/maps.htm.
https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1
https://www.intelligentautomation.network/decision-ai/articles/a-basic-guide-to-predictive-analytics
https://towardsdatascience.com/cluster-then-predict-for-classification-tasks-142fdfdc87d6
https://www.fisheries.noaa.gov/foss/f?p=215:200:13846825675961:::::
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8776938/

Visualizations:

All relevant visualizations can be found in the Google Collab folders located in our GitHub repository.