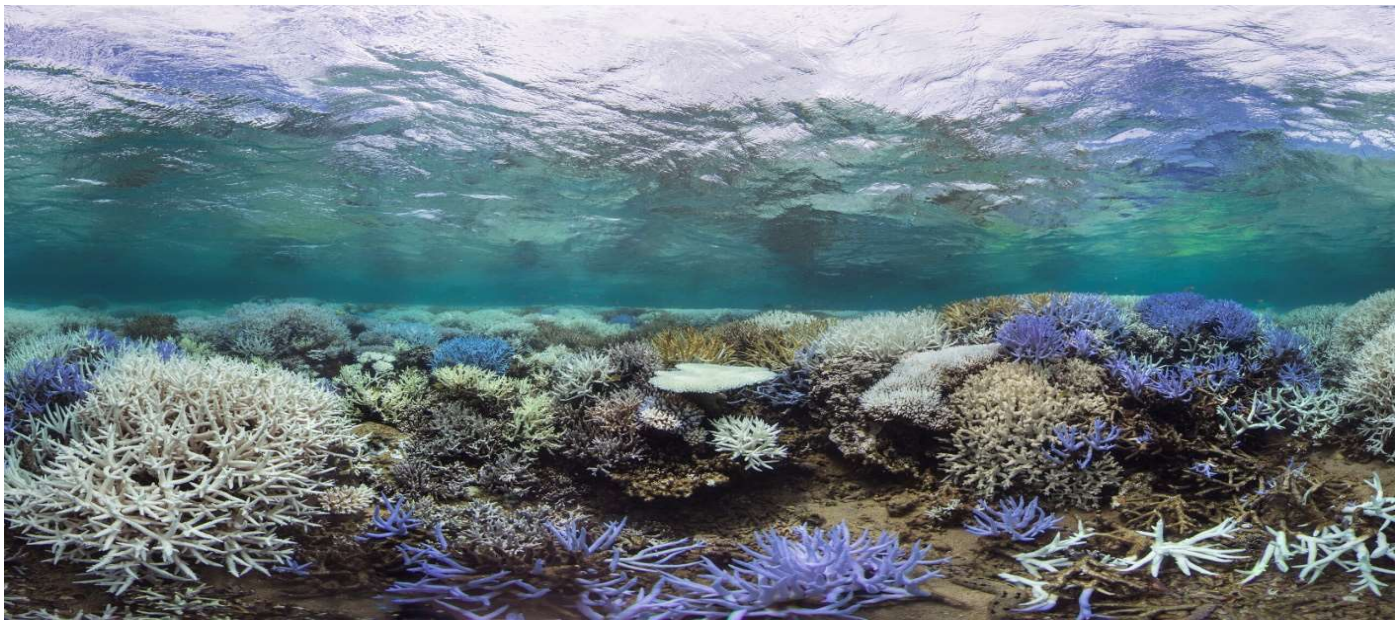
Predicting Coral Bleaching Events Through Environmental Data Analysis



By:

Mathurin, Gregory

Ko, Graeme

Al-Saidi, Ray

Table of Contents

Introduction

Coral reefs, often referred to as the "rainforests of the sea," are vital to the health of the world's oceans and to human societies. These ecosystems are incredibly diverse, supporting more species per area than any other marine environment. This includes about 4,000 species of fish, 800 species of hard corals, and potentially millions of undiscovered species. Beyond their biological importance, coral reefs provide significant economic benefits. They are crucial for commercial and subsistence fisheries, contribute to jobs and businesses through tourism and recreation, and protect coastlines from the energy of waves, storms, and floods. The commercial value of U.S. fisheries from coral reefs alone is over $100 million, while the global value of coral reefs is estimated at £6 trillion annually, due to their contribution to fishing, tourism industries, and coastal protection (Natural History Museum Accessed 10 Feb. 2024).

Coral reefs also have a unique symbiotic relationship with algae called zooxanthellae, which live within the coral and are essential for their survival. This relationship is sensitive to changes in the environment, particularly to temperature, which can lead to coral bleaching when the water is too warm. Coral reefs are facing threats from pollution, disease, habitat destruction, rising ocean temperatures, and ocean acidification, making their conservation a priority for global biodiversity and human economies (US EPA, OA. U.S. Environmental Protection Agency. 20 Mar. 2013).

Preserving coral reefs is not only about protecting a beautiful and vital component of the marine ecosystem; it is also about safeguarding the livelihoods of millions of people and the coastal communities that depend on them. Coral reefs act as natural barriers, protecting shorelines from erosion and storm surges, and are a source of food and new medicines. Their decline has profound implications for marine biodiversity and for people who depend on them for food, income, and protection against natural disasters (US EPA, OA. U.S. Environmental Protection Agency. 20 Mar. 2013).

Purpose

Exploring coral reefs serves several essential purposes that directly impact the health of marine habitats and the balance of our planet's ecosystems. Biodiversity assessments and ecological research help us understand the intricate dynamics of coral reef environments, which are crucial for sustaining diverse marine life and supporting the production of oxygen vital for human populations. Monitoring efforts and studies on climate change impacts provide insights into the vulnerability of coral reefs to environmental stressors, such as coral bleaching, which threatens the health and resilience of these ecosystems.

By identifying key factors leading to bleaching and implementing conservation strategies informed by research and statistical analysis, we can work towards creating better habitats and safeguarding the essential functions of coral reefs. Additionally, economic valuations highlight the significance of coral

reefs in supporting fisheries, tourism, and coastal protection, underscoring the importance of conservation efforts for both ecological and socioeconomic reasons. Engaging communities and stakeholders in coral reef research and conservation initiatives would further strengthen our collective ability to preserve these invaluable ecosystems for future generations.

Objectives

1. To investigate the fundamental environmental factors influencing coral bleaching to discern the primary drivers behind these events and where they occur.

2. Subsequently, to construct a robust predictive model for coral bleaching occurrences, utilizing comprehensive environmental datasets to forecast the likelihood of bleaching events.

3. Finally, to rigorously assess the accuracy and reliability of the predictive model in anticipating bleaching events, thereby gauging its practical utility for informing proactive conservation strategies within marine ecosystems.

The Dataset

The dataset, funded by the U.S. National Science Foundation and sourced from the Biological and Chemical Oceanography Data Management Office, represents a comprehensive collection of global bleaching environmental data. It encompasses 41,361 entries across 62 columns, featuring variables such as latitude, longitude, ocean name, various sea surface temperature metrics (mean, max, frequency, etc.), degree heating weeks, and comments on bleaching events. The data spans several years, covering multiple geographic locations and oceanic realms, and the version used for this analysis was released in 2022, specifically focusing on measurements from 1980 to 2020. Initially, the dataset contained 62 variables; however, many were deemed irrelevant to our analysis due to their lack of relevance or high collinearity and were thus excluded. For further details on the dataset's variables and structure, the data dictionary is found under Appendix C.

Methodology

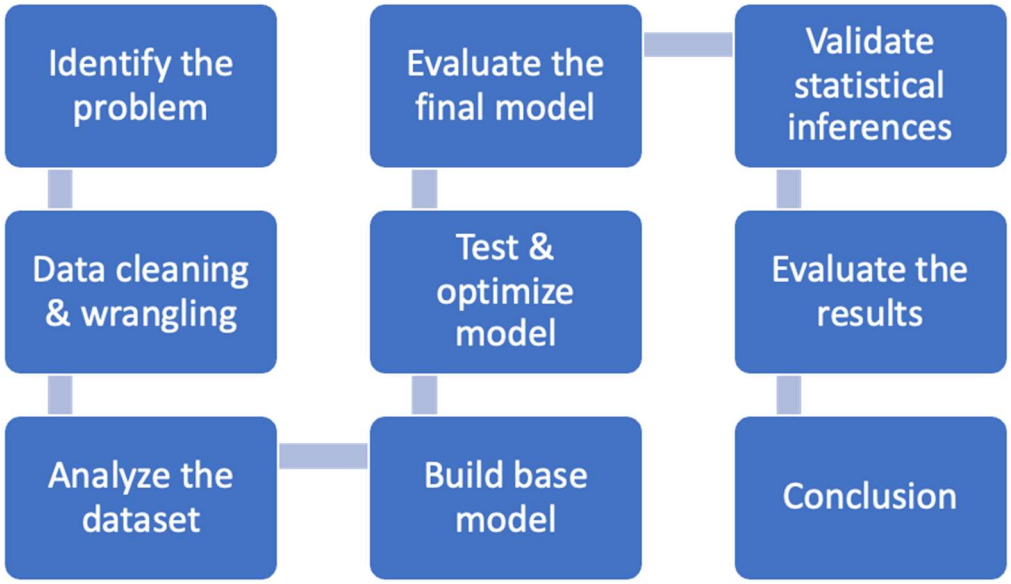


Figure 1: Methodology Chart

1. Data Preprocessing: Clean the dataset to handle missing values, outliers, and categorize textual data for analysis. Also, we will remove unnecessary columns based on multicollinearity, possible independence issues, and significance to our work.

2. Exploratory Data Analysis: Perform statistical analysis to understand the distribution of key variables and their relationship with coral bleaching events.

3. Feature Selection: Identify the most relevant environmental factors that contribute to coral bleaching using correlation analysis and feature importance techniques.

4. Model Development: Develop a model (Multiple Linear Regression, linear discriminant analysis, logistic regression) to predict coral bleaching events based on environmental factors.

5. Model Evaluation: Use metrics such as accuracy, precision, recall, and AUC-ROC curve to evaluate model performance.

6. Interpretation and Recommendations: Analyze the model's findings to interpret the environmental conditions most predictive of bleaching events and provide recommendations for conservation efforts.

Data Exploration

Initial visualizations to explore the data were created using Tableau, providing insights into the various dimensions of the dataset, including temporal trends and spatial distribution of bleaching events. These tools combined offered a powerful approach for data cleaning, manipulation, and visualization, enabling a comprehensive analysis of the environmental factors contributing to coral bleaching. We decided to conduct some data exploration prior to data analysis to explore the relationships between variables, identify patterns, and gain insights into the underlying structure of the data. This preliminary step helped us understand the characteristics of the dataset, uncover potential issues such as missing values or outliers, and inform our subsequent data analysis strategies. Our initial insights were as follows:

Insight #1

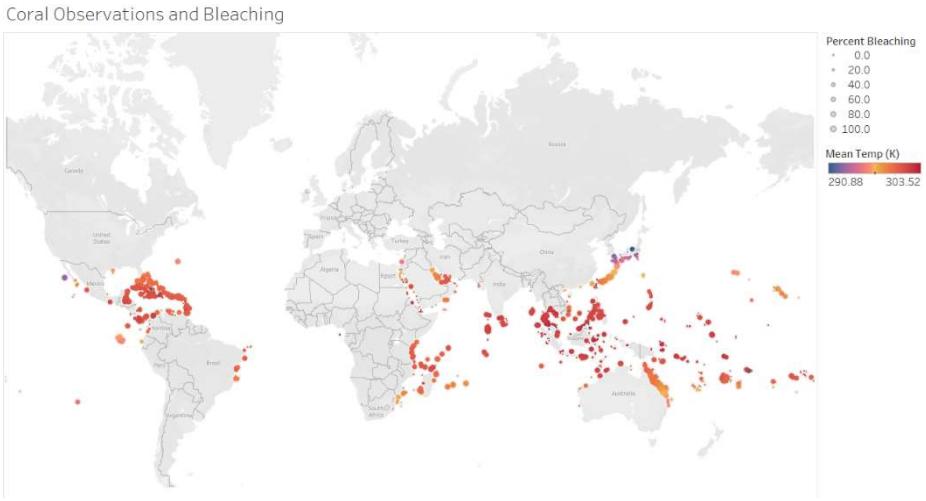


Figure 2: Bleaching Percentage/Mean Temperature

We tried plotting every observation where the colour depicts the mean temperature of that site measured in Kelvin, and percent bleaching corresponds to the size of the data point. We can see that many of the significant bleaching sites also have a high mean temperature, though other sites with a lower temperature also experienced bleaching, perhaps indicating that temperature alone is not the only factor to consider.

Insight #2

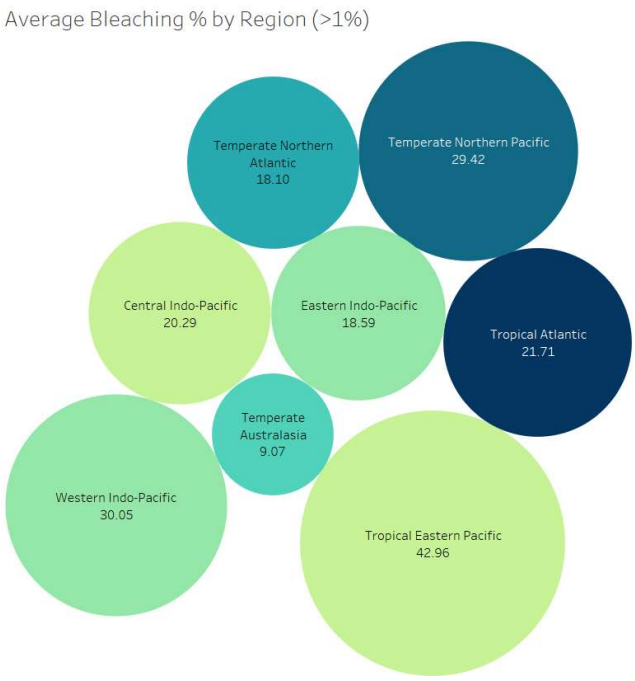


Figure 3: Average Bleaching by Region

Here we split the oceans into different regions and plotted their average percent bleaching values. We decided to look only at the values with >%1 bleaching when calculating these averages to focus only on the severity of coral reefs either destroyed or at risk. We can see from the graph that the Tropical Eastern Pacific region had the highest average severity bleaching at 43% while Temperate Australasia had the least at 9%.

Insight #3

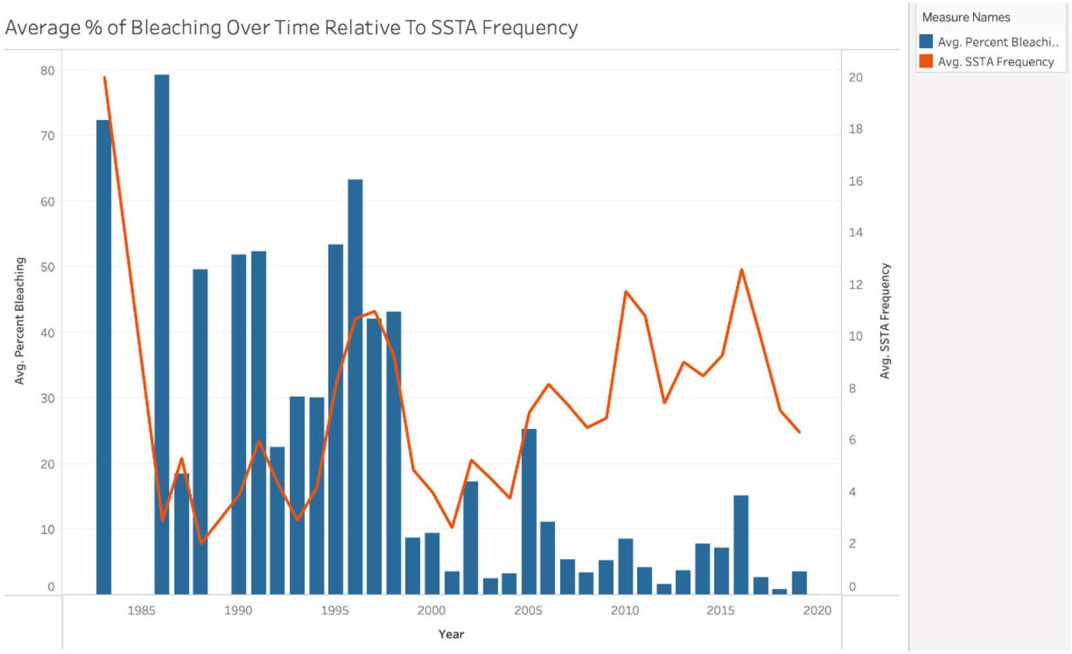


Figure 4 Bleaching/SSTA vs Year

Based on the visualization above, focusing on data from 2000 onwards which comprises over 90% of our dataset, we observe years with significant temperature variation, notably 2005, 2010, and 2015, as evidenced by peaks in Sea Surface Temperature Anomaly (SSTA) frequency, correspond to higher-than-average coral bleaching percentages which is interesting to keep in mind during our analysis.

Data Cleaning & Wrangling

Data cleaning and preparation were conducted using Python's pandas library. We initially had 62 different variables. We then restructured the dataset by removing variables that did not contribute to the regression analysis, contained a high proportion of null values, had potential independence issues, or exhibited high collinearity with other variables. We were able to narrow it down to 15 key variables of interest, of which we then removed any rows containing null values.

The following variables exemplify which were removed:

Sample_Comments	State_Island_Province_Name
Bleaching_Comments	Country_Name
Site_Comments	Latitude_Degrees
TSA_DHW_Standard_Deviation	Longitude_Degrees
Date	City_Town_Name
TSA_Frequency_Standard_Deviation	Reef_ID
TSA_Standard_Deviation	Sample_ID
SSTA_DHW_Standard_Deviation	Site_ID
SSTA_Frequency_Standard_Deviation	Date_Day
SSTA_Standard_Deviation	Date_Month
Temperature_Kelvin_Standard_Deviation	Date_Year
Site_Name	SSTA_Mean

While the 15 variables we kept included:

Ocean_Name	Distance_to_Shore
Exposure	Turbidity
Cyclone_Frequency	Depth_m
Percent_Bleaching	Temperature_Kelvin
Temperature_Mean	Temperature_Minimum
Temperature_Maximum	Windspeed
SSTA_Maximum	SSTA_Frequency
SSTA	

Analysis

Linear Regression:

We first began by building a multiple linear regression model using the 15 key variables of interest. After removing variables which contained high correlation and multicollinearity, we then used the ols_best_subset() method in R to select the variables for our base model.

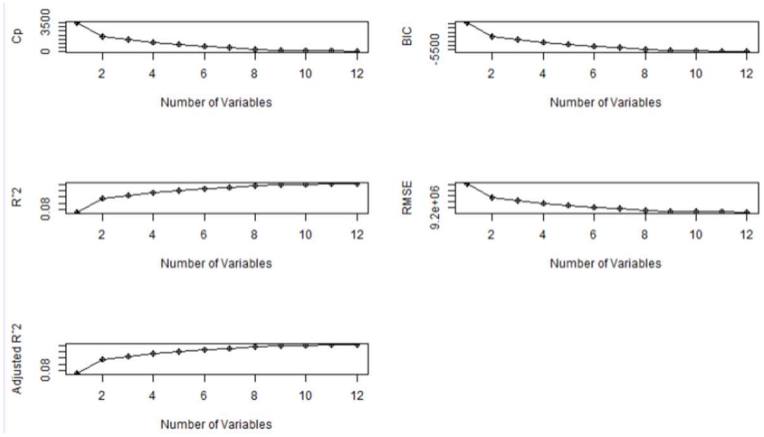


Figure 5: OLS Subset Output

We chose a variable count of 10, as it had a low Mallow's Cp criterion while still being close to the number of predictors + 1, perhaps indicating a good value without incorporating too much bias into the model. Furthermore, it still had a somewhat low AIC and high R-squared value compared to lower variable models and similar scores compared to higher variable ones.

```
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -32.65171   22.68551  -1.439  0.15007
Turbidity        -21.45922    1.75407 -12.234  < 2e-16 ***
Cyclone_Frequency  -0.06687    0.01375  -4.863 1.16e-06 ***
Depth_m             0.31586    0.02339  13.504  < 2e-16 ***
Temperature_Kelvin  1.32067    0.06061  21.790  < 2e-16 ***
Temperature_Mean   -1.21976    0.08013 -15.222  < 2e-16 ***
Windspeed           0.34476    0.05201   6.629 3.44e-11 ***
SSTA_Maximum        0.28101    0.08804   3.192  0.00141 **
SSTA_Frequency      0.61346    0.01588  38.620  < 2e-16 ***
Ocean_NameAtlantic 10.26972    0.95785  10.722  < 2e-16 ***
Ocean_NameIndian    7.82562    1.00541   7.784 7.26e-15 ***
Ocean_NamePacific   1.77202    0.93984   1.885  0.05938 .
Ocean_NameRed Sea  -5.83146    1.09774  -5.312 1.09e-07 ***
ExposureSheltered  -1.64024    0.21964  -7.468 8.36e-14 ***
ExposureSometimes   5.62740    0.35975  15.642  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.63 on 32699 degrees of freedom
Multiple R-squared:  0.1622,    Adjusted R-squared:  0.1618
F-statistic: 452.1 on 14 and 32699 DF,  p-value: < 2.2e-16
```

We can see that turbidity had the greatest effect on the linear model. The base model yielded a rather low R-squared adjusted value of 0.16 so we decided to try an interaction model as well.

Significant Interactions:	
Turbidity:Exposure	Windspeed:Exposure
Turbidity:Ocean_Name	SSTA_Maximum:Ocean_Name
Depth_m:Ocean_Name	SSTA_Maximum:Exposure
Depth_m:Exposure	SSTA_Frequency:Ocean_Name
Temperature_Kelvin:Ocean_Name	SSTA_Frequency:Exposure
Temperature_Mean:Exposure	Ocean_Name:Exposure

However, this interaction model, while improved, still yielded a low R-squared adjusted of 0.23. Higher order terms were considered to further improve the model, but scatterplots done on the variables did not reveal any clear non-linear relationships to investigate. Therefore, we kept the interaction model and performed further tests on it. We looked at high leverage and outlier points within our dataset which we then removed, though it did not improve the model.

Figure 5: Fitted Vs. Residuals

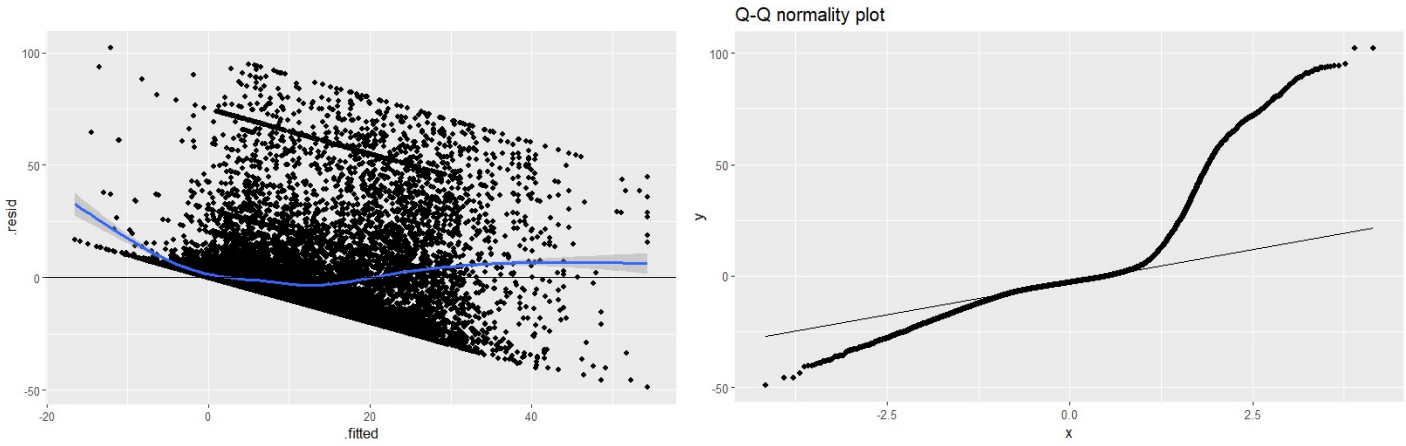


Figure 6: Q-Q Plot

We then plotted the residuals, showing that it did not pass the homoscedasticity or normality assumptions. We did not attempt any transformations such as box-cox transformations to improve this given the already poor fit. Instead, we decided to explore other models.

Linear Discriminant Analysis (LDA):

After conducting multiple linear regression as part of our objectives, we aimed to explore an alternative approach for predicting coral bleaching events. Consequently, we opted to address this challenge from a classification perspective. The procedure and results were as follows:

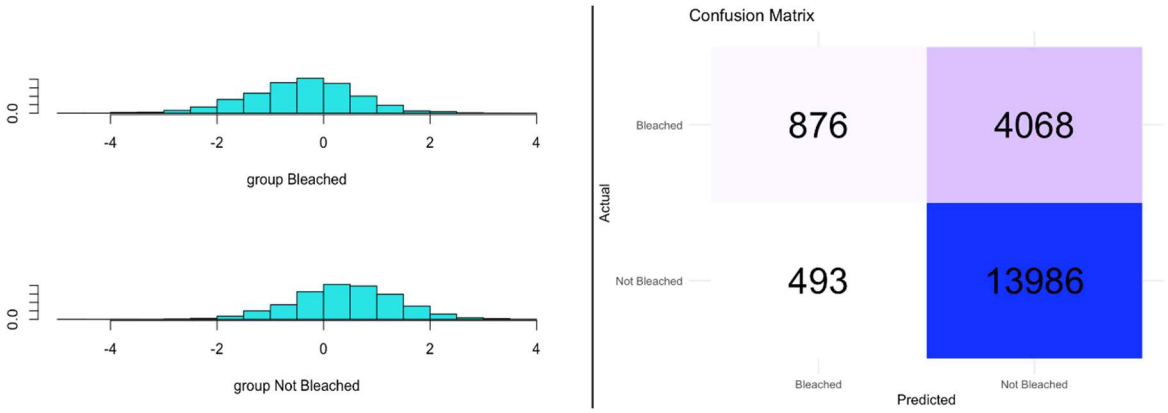
1. The first step in the process was, using the same 15 key variables of interest as multiple linear regression, to change the variable “Percent Bleaching” into a categorical variable for our prediction. We reclassified this response variable into a newly created column titled "Bleached." This reclassification followed a simplified version of the categorization specified in the scientific journal article, "A New, High-Resolution Global Mass Coral Bleaching Database." The categorization details were as follows:
   - 0 = No bleaching
   - 1 = Mild bleaching (0-10%)
   - 2 = Moderate bleaching (11-50%)
   - 3 = Severe bleaching (>50%)

   For simplicity, we wanted to initially construct the LDA to predict two binary outcomes for the coral events and as a result we developed a new criterion as follows for the applicable “Bleached” column:

   - 0 = Bleaching of 10% or less, labeled as "not bleached."
   - 1 = Bleaching greater than 10%, labeled as "bleached."

2. Using the newly created class/categorial variable “Bleached”, we wanted to split the data to allow for a 75%/25% split for the training and test sets, respectively, while also concurrently applying stratified sampling. Stratified sampling would allow us to prevent bias and ensure that our LDA model is trained and evaluated on a better representative sample of the entire population.

3. Following the model building, the misclassification rate obtained by the LDA model predicting the coral bleaching events from the test data was ~15.6% which better than initially expected given the complexity of our objective and the results are further elaborated by the confusion matrix below.



As for the distribution of discriminant scores, it was noted that the LDA model provided some clear distinctions in the discriminant scores between each class, however, some clear overlaps were identified and reflected in the misclassification rate obtained.

4. To ensure the reliability of the model prior to any optimization procedures, we wanted to check the related assumptions. The following methods were applied:
    a. Independence Assumption: The LDA requires that knowing the value of one predictor variable should not provide any information about the values of other predictor variables within the same class. This assumption was tested through rationalizations and visualizations and was further deemed to hold true.
    b. Normality Assumption: the LDA model required that the multivariate normality assumption held true throughout the process. This assumption was tested using the Energy Test.
    c. Homoscedasticity Assumption: The LDA model required the assumption of equal covariance among K classes. This assumption was tested using the Box's M Test.

Assumptions	Method	Result
Normality	Energy Test	p-value < 2.2e-16
Homoscedasticity	Box's M Test	p-value < 2.2e-16

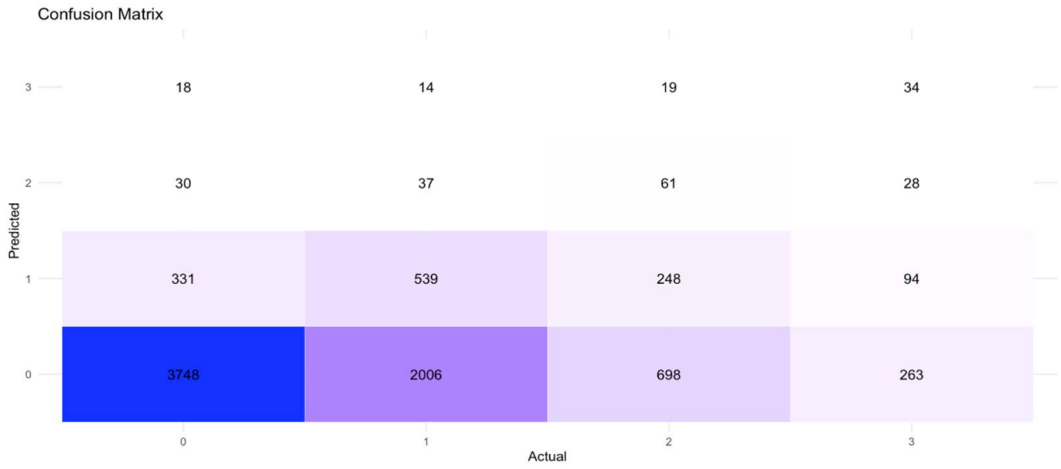
Given the extremely low values obtained on both tests, which ultimately rejects our null hypothesis, we concluded that our assumptions on normality and homoscedasticity did not hold true.

We did not attempt to optimize or modify the LDA model further as we concluded that it was not viable to rely on given that it did not pass the assumption tests. As a result, we decided to try alternative classification methods to get more reliable results.

Logistic Regression:

In our evaluation of the model using logistic regression, we initially explored multinomial regression. For this analysis, we reclassified the response variable "bleached percentage" into a newly created column titled "Bleached_multi." This reclassification followed the same 4 categories listed in the LDA section.

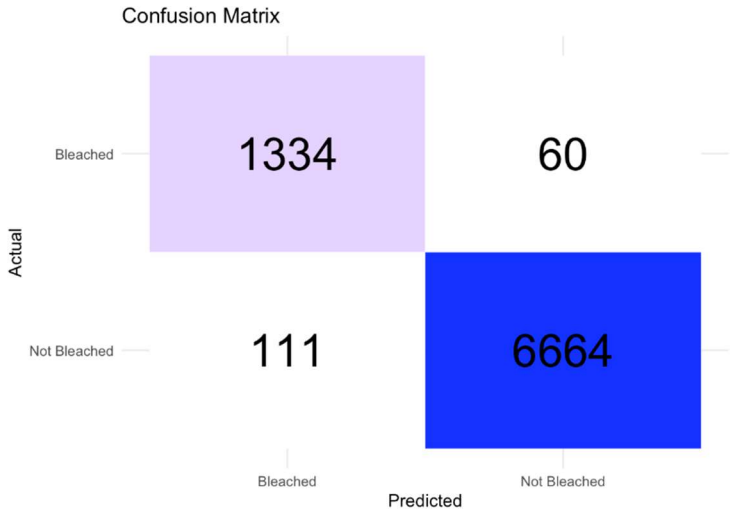
To achieve a balanced representation of each outcome and enhance the accuracy of our model, we utilized similar sampling scheme as LDA, adhering to a 75/25 split. Despite these methodological efforts, the approach resulted in a classification rate of 53.64%. This outcome highlights the inherent difficulties in accurately modeling the complex gradations of coral bleaching through multinomial logistic regression. Below is the resulting confusion matrix.



We shifted our focus to binary logistic regression. This decision was based on the hypothesis that a simpler model might improve predictive accuracy. We continued to use the binary classification system established in LDA, defining '0' as 10% or less bleaching ("not bleached") and '1' as more than 10% bleaching ("bleached").

Originally, our binary model incorporated 27 variables, achieving a classification rate of 84%. However, after conducting a Variance Inflation Factor (VIF) test to assess multicollinearity, we pruned the model by removing several variables deemed redundant or overly influential.

We ended up using the same variables employed in linear regression. We achieved an 82% classification rate. Although this represents a slight decline in performance, it likely increases the model's robustness and interpretability. Below is the confusion matrix:



The confusion matrix reveals a high number of true positives and true negatives, indicating the model's strength in accurately predicting both bleached and not bleached outcomes. The relatively low false positive and false negative rates suggest that the model is reliable for practical applications, such as

monitoring coral reef health and predicting future bleaching events. The model incorporated these 11 variables and was refined through five iterations of Fisher Scoring. The findings are summarized in the table below:

Variable	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-52.79006	5.44339	-9.698	< 0.0000000000000002
Distance_to_Shore	0	0	1.918	0.0551
Turbidity	-6.25631	0.44407	-14.089	< 0.0000000000000002
Cyclone_Frequency	0.01117	0.00229	4.882	1.05E-06
Depth_m	0.07846	0.00404	19.404	< 0.0000000000000002
Temperature_Kelvin	0.33244	0.01578	21.072	< 0.0000000000000002
Temperature_Mean	-0.37322	0.01949	-19.153	< 0.0000000000000002
Temperature_Maximum	0.20331	0.02593	7.841	4.47E-15
Windspeed	0.08491	0.0097	8.758	< 0.0000000000000002
SSTA	-0.12333	0.02685	-4.594	4.35E-06
SSTA_Maximum	-0.15079	0.02434	-6.194	5.87E-10
SSTA_Frequency	0.05853	0.00285	20.502	< 0.0000000000000002

The significance of Turbidity as a predictor is particularly noteworthy. Its negative coefficient suggests that higher turbidity levels are associated with a decreased likelihood of coral bleaching. This relationship highlights the protective effect of water turbidity against bleaching, potentially by reducing the intensity of light reaching the corals and thereby mitigating the thermal stress caused by elevated temperatures.

However, the insignificance of Distance to Shore as a predictor challenges common sense. It implies that the direct impact of proximity to land on bleaching events may be less critical than previously thought. This finding prompts a re-evaluation of the spatial factors considered in coral health assessments and suggests that other, more complex environmental interactions may play a more crucial role.

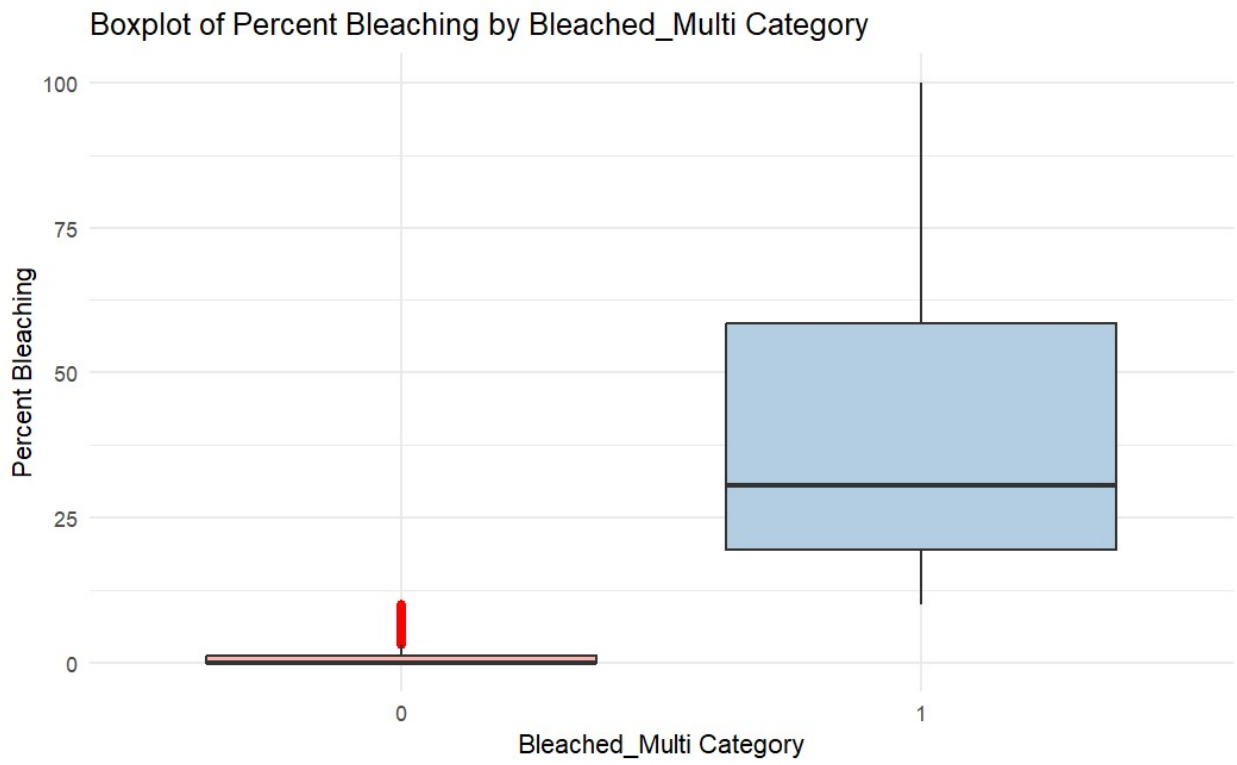
The variables related to temperature — Temperature (Kelvin), Temperature Mean, and Temperature Maximum — all demonstrate significant effects, with their coefficients indicating the critical role of thermal conditions in coral bleaching. The positive coefficient for Temperature (Kelvin) and Temperature Maximum suggests that as water temperatures rise, so does the likelihood of bleaching. The negative coefficient for Temperature Mean, however, suggests a delicate relationship where average conditions might mitigate some risks, through acclimatization effects or the presence of thermally tolerant coral species.

The significance of SSTA (Sea Surface Temperature Anomaly) variables, including SSTA Maximum and SSTA Frequency, further emphasizes the impact of temperature variations on coral health. These findings align with existing research, which identifies thermal stress as a primary driver of coral bleaching events globally.

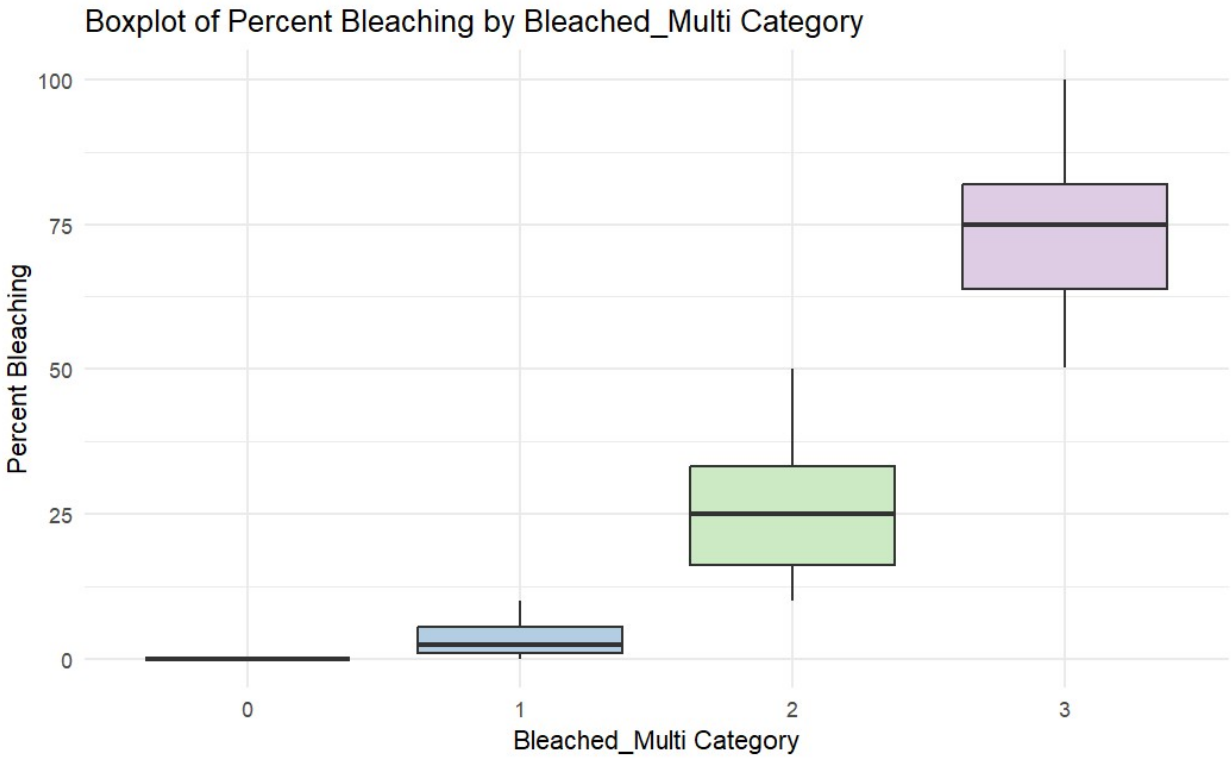
Lastly, the Cyclone Frequency's positive coefficient introduces an interesting perspective on the role of cyclonic events influencing coral ecosystems. While cyclones can cause direct physical damage to reefs, their associated cooling effects and nutrient mixing might have a complex impact on coral health and bleaching responses.

To explain the comparative effectiveness of our logistic regression approaches, we graphically represented the classification schemes for the response variable. This visual comparison clarified the distinctions between categories, revealing that the binary logistic regression framework provided broader categorization boundaries. This adaptability underpins the superior classification rate of the binary model.

Logistic Regression Boxplot:



The multinomial logistic regression accommodates multiple categories, which results in narrower categorization boundaries. Consequently, there is an increased likelihood of misclassification. Relevant box plots can be seen below.



Conclusion

Our analysis revealed that among the models tested, logistic regression demonstrated the highest classification accuracy while adhering to the independence condition. This conclusion is supported by our model results, which are as follows:

- Linear Regression exhibited an R^2 value of 0.233, indicating a modest fit to the data, however failed to meet the assumptions for the model.
- Linear Discriminant Analysis had a misclassification rate of 15.62%, suggesting a relatively high level of accuracy in classification but failed to meet the required conditions under LDA.
- Logistic Regression, however, had a misclassification rate of 17.02%, which, despite being slightly higher than that of Linear Discriminant Analysis, was deemed the most suitable model due to its adherence to the independence condition.

Additionally, our observations indicated that the coral reefs in the Tropical Eastern Pacific region experienced the highest levels of bleaching. This finding underscores the urgent need for targeted environmental protection measures in this area.

Based on these insights, we recommend the following actions to enhance environmental monitoring and management:

- Deploy advanced technologies and establish real-time monitoring systems. These systems should be capable of issuing instant alerts for significant changes in key environmental indicators such as turbidity, temperature, or Sea Surface Temperature Anomalies (SSTAs) that exceed established

safety thresholds. This proactive approach will allow for timely interventions to mitigate potential environmental threats.

- Initiate comprehensive and interdisciplinary studies aimed at integrating data on human population dynamics and pollution levels into our current models. This approach will provide a more holistic understanding of the environmental impact of human activities and improve the accuracy of our predictions and the effectiveness of our mitigation strategies.
- Focus conservation efforts on the Tropical Eastern Pacific coral reefs, developing specific strategies to combat the high rates of bleaching observed in this region. This may include measures to reduce local sources of pollution, protect against overfishing, and mitigate the impacts of climate change.

These steps are essential for developing a more resilient and responsive environmental monitoring system, ultimately contributing to the preservation of our natural resources and the well-being of our communities.

References

Home | Natural History Museum. https://www.nhm.ac.uk. Accessed 10 Feb. 2024.

How Much Oxygen Comes from the Ocean? https://oceanservice.noaa.gov/facts/ocean-oxygen.html. Accessed 10 Feb. 2024.

Jiamwattanapong, K., Ingadapa, N., & Plubin, B. (2021). On testing homogeneity of covariance matrices with box's m and the approximate tests for multivariate data. European Journal of Applied Sciences, 9(5), 426–436. https://doi.org/10.14738/aivp.95.11115

US EPA, OA. U.S. Environmental Protection Agency. 20 Mar. 2013, https://www.epa.gov/home.

Zach. (2020, October 5). How to perform multivariate normality tests in r. Statology. https://www.statology.org/multivariate-normality-test-r/

Appendix A: Python Code

```
1 import pandas as pd
2 import numpy as np
```

```
1 coral = pd.read_csv("global_bleaching_environmental.csv")
2 coral.replace('nd', np.nan, inplace=True)
3 np.shape(coral)
4 coral.dropna(subset=["Percent_Bleaching"],inplace=True)
5 coral.head()
```

```
<ipython-input-2-28ea534af44d>:1: DtypeWarning: Columns (13,15,24) have mixed types. Specify dtype o coral =
  pd.read_csv("global_bleaching_environmental.csv")
```

	Site_ID	Sample_ID	Data_Source	Latitude_Degrees	Longitude_Degrees	Ocean_Name	Reef_ID	Realm_
0	2501	10324336	Donner	23.163	-82.5260	Atlantic	NaN	Tro Atl
1	3467	10324754	Donner	-17.575	-149.7833	Pacific	NaN	Eas Indo-Pa
2	1794	10323866	Donner	18.369	-64.5640	Atlantic	NaN	Tro Atl
3	8647	10328028	Donner	17.760	-64.5680	Atlantic	NaN	Tro Atl
4	8648	10328029	Donner	17.769	-64.5830	Atlantic	NaN	Tro Atl

5 rows × 62 columns

```
1 np.shape(coral)
```

```
(34515, 62)
```

```
 1 useless =                    ["Data_Source","Sample_Comments","Bleaching_Comments","Site_Comments",\
 2    "TSA_DHW_Standard_Deviation","Date","TSA_Frequency_Standard_Deviation","TSA_Standard_Devia
 3       "SSTA_DHW_Standard_Deviation","SSTA_Frequency_Standard_Deviation","SSTA_Standard_Deviation"
 4             "Temperature_Kelvin_Standard_Deviation","Site_Name","State_Island_Province_Name",
 5             "Country_Name","Latitude_Degrees","Longitude_Degrees","City_Town_Name","Reef_ID",
 6                        "Sample_ID","Site_ID","Date_Day","Date_Month","Date_Year","SSTA_Mean"]
 7
 8 #SSTA_Mean is all 0 in the dataset 9
10 coral.drop(columns=useless, inplace=True)
11 np.shape(coral)
```

```
(34515, 37)
```

```
1 coral.isna().sum()
```

```
Ocean_Name              0
Realm_Name              0
Ecoregion_Name          3
Distance_to_Shore       2
Exposure                0
Turbidity               6
Cyclone_Frequency       0
Depth_m              1681
Substrate_Name      12047
Percent_Cover       11842
Bleaching_Level     11984
Percent_Bleaching       0
ClimSST                95
Temperature_Kelvin    122
Temperature_Mean      106
Temperature_Minimum   106
Temperature_Maximum   106
Windspeed             111
SSTA                  122
SSTA_Minimum          142
SSTA_Maximum          106
SSTA_Frequency        122
SSTA_FrequencyMax     106
SSTA_FrequencyMean    106
SSTA_DHW              122
SSTA_DHWMax           106
SSTA_DHWMean          106
TSA                   122
TSA_Minimum           106
TSA_Maximum           106
TSA_Mean              106
TSA_Frequency         122
TSA_FrequencyMax      106
TSA_FrequencyMean     106
TSA_DHW               122
TSA_DHWMax            106
TSA_DHWMean           106
dtype: int64
```

```
remove_maybe = ["Substrate_Name","Percent_Cover","Bleaching_Level","Ecoregion_Name","Realm_Name","Cli
#Ecoregion_Name has 100 unique values and Realm_Name (with 8 unique values) is more precise than Ocea
 coral.drop(columns=remove_maybe,  inplace=True)

remove_DHW = ["SSTA_DHW","SSTA_DHWMax","SSTA_DHWMean","TSA_DHW","TSA_DHWMax","TSA_DHWMean"]
#Remove DHW for simplicity and perhaps dependence with the frequency variables
 coral.drop(columns=remove_DHW,  inplace=True)

remove_TSA = ["TSA","TSA_Minimum","TSA_Maximum","TSA_Mean","TSA_Frequency","TSA_FrequencyMax","TSA_Fr
#Remove TSA for simplicity and dependence
 coral.drop(columns=remove_TSA,  inplace=True)

 remove_SSTA_unecessary  =  ["SSTA_FrequencyMax","SSTA_FrequencyMean","SSTA_Minimum"]
#Remove extra SSTA for simplicity and dependence
#Take out things like SSTA_FrequencyMax and SSTA_FrequencyMean to instead capture only year-long data
# The effects of recent climate change more rather than looking at decades-long intervals.
 coral.drop(columns=remove_SSTA_unecessary,  inplace=True)

np.shape(coral)
```

```
(34515, 32)
```

```
1 coral.dropna(inplace=True)
```

```
1 coral.head()
```

	Realm_Name	Distance_to_Shore	Exposure	Turbidity	Cyclone_Frequency	Depth_m	Perce
0	Atlantic Tropical	8519.23	Exposed	0.0287	49.90	10	
1	Indo-Pacific Eastern	1431.62	Exposed	0.0262	51.20	14	
2	Atlantic Tropical	182.33	Exposed	0.0429	61.52	7	
3	Atlantic Tropical	313.13	Exposed	0.0424	65.39	9.02	
4	Atlantic Tropical	792.0	Exposed	0.0424	65.39	12.50	

5 rows × 32 columns

```
1 np.shape(coral)
```

```
(32678, 32)
```

```
1 coral.nunique()
```

```
Realm_Name              8
Distance_to_Shore    9943
Exposure                3
Turbidity            2020
Cyclone_Frequency    1299
Depth_m               468
Percent_Bleaching    2364
ClimSST               916
Temperature_Kelvin   1170
Temperature_Mean      739
Temperature_Minimum   887
Temperature_Maximum   587
Windspeed              18
SSTA                  621
SSTA_Minimum          348
SSTA_Maximum          505
SSTA_Frequency        372
SSTA_FrequencyMax     272
SSTA_FrequencyMean    141
SSTA_DHW             1519
SSTA_DHWMax          1767
SSTA_DHWMean          445
TSA                  1081
TSA_Minimum           840
TSA_Maximum           399
TSA_Mean              559
TSA_Frequency         184
TSA_FrequencyMax      201
TSA_FrequencyMean      67
```

```
TSA_DHW        960
TSA_DHWMax    1442
TSA_DHWMean    214
dtype: int64
```

```
1 coral.to_csv("Clean_Coral_Data.csv",  header=True,index=False)
```

Appendix B: R Code

Please check attached file

Appendix C: Data Dictionary

Parameter	Description	Units
Site_ID	Unique identifier for each site	unitless
Sample_ID	Unique identifier for each sampling event	unitless
Data_Source	Source of data set	unitless
Latitude_Degrees	Latitude coordinates (positive vaues = North; negative values = South)	degrees North
Longitude_Degrees	Longitude coordinates (positive values = East; negative values = West)	degrees East
Ocean_Name	The ocean in which the sampling took place	unitless
Reef_ID	Unique identifier from Reef Check data	unitless
Realm_Name	Identification of realm as defined by the Marine Ecoregions of the World (MEOW) Spalding et al. 2007	unitless
Ecoregion_Name	Identification of the Ecoregions (150) as defined by Veron et al	unitless
Country_Name	The country where sampling took place	unitless
State_Island_Province_Name	The state, territory (e.g., Guam) or island group (e.g., Hawaiian Islands) where sampling took place	unitless
City_Town_Name	The region, city, or nearest town, where sampling took place	unitless
Site_Name	The accepted name of the site or the name given by the team that sampled the reef	unitless

Distance_to_Shore	The distance of the sampling site from the nearest land	meters (m)
Exposure	The site's exposure to fetch. Site was considered exposed if it had >20 km of fetch, if there were strong seasonal winds, or if the site faced the prevailing winds. Otherwise, the site was considered sheltered or "sometimes". "Sometimes" refers to a few sites with a >20 km fetch through a narrow geographic window, and therefore we considered that the site was potentially exposed during cyclone seasons.	unitless
Turbidity	Kd490 with a 100-km buffer. Turbidity was considered to be positively related to the diffuse attenuation coefficient of light at the 490 nm wavelength (Kd490), or the rate at which light at 490 nm is attenuated with depth. For example, a Kd490 value of 0.1 m?1 means that light intensity is reduced by one natural-log value within 10 m of water. High values of Kd490, therefore, represent high attenuation and hence high turbidity.	reciprocal meters (m-1)
Cyclone_Frequency	number of cyclone events from 1964 to 2014	unitless
Date_Day	the day of the sampling event	unitless
Date_Month	the month of sampling event	unitless
Date_Year	the year of sampling event	unitless
Depth_m	depth of sampling site	meters (m)

Substrate_Name	type of substrate from Reef Check data	unitless
Percent_Cover	average cover value (percent)	percent
Bleaching_Level	Reef Check data, coral population or coral colony	unitless
Percent_Bleaching	An average of four transect segments (Reef Check) or average of a bleaching code	percent
ClimSST	Climatological sea surface temperature (SST) based on weekly SSTs for the study time frame, created using a harmonics approach	degrees Celsius
Temperature_Kelvin	Temperature in Kelvin	Kelvin
Temperature_Mean	Mean Temperature	degrees Celsius
Temperature_Minimum	Minimum Temperature	degrees Celsius
Temperature_Maximum	Maximum Temperature	degrees Celsius
Temperature_Kelvin_Standard_Deviation	Standard deviation of temperature	Kelvin
Windspeed	Windspeed	meters per hour
SSTA	Sea Surface Temperature Anomaly: weekly SST minus weekly climatological SST	degrees Celsius
SSTA_Standard_Deviation	The Standard Deviation of weekly SST Anomalies over the entire time period	degrees Celsius
SSTA_Mean	The mean SSTA over the entire time period	degrees Celsius
SSTA_Minimum	The minimum SSTA over the entire time period	degrees Celsius

SSTA_Maximum	The maximum SSTA over the entire time period	degrees Celsius
SSTA_Frequency	Sea Surface Temperature Anomaly Frequency: number of times over the previous 52 weeks that SSTA >=1 degree C	SSTA per time period
SSTA_Frequency_Standard_Deviation	The standard deviation of SSTA_Frequency over the entire time period	SSTA per time period
SSTA_FrequencyMax	The maximum SSTA_Frequency over the entire time period	SSTA per time period
SSTA_FrequencyMean	The mean SSTA_Frequency over the entire time period	SSTA per time period
SSTA_DHW	Sea Surface Temperature Degree Heating Weeks: sum of previous 12 weeks when SSTA>=1 degree C	weeks
SSTA_DHW_Standard_Deviation	The standard deviation SSTA_DHW over the entire time period	weeks
SSTA_DHWMax	The maximum SSTA_DHW over the entire time period	weeks
SSTA_DHWMean	The mean SSTA_DHW over the entire time period	weeks
TSA	Thermal Stress Anomaly: Weekly sea surface temperature minus the maximum of weekly climatological sea surface temperature	degrees Celsius
TSA_Standard_Deviation	The standard deviation of TSA over the entire time period	degrees Celsius
TSA_Minimum	The minimum TSA over the entire time period	degrees Celsius

TSA_Maximum	The maximum TSA over the entire time period	degrees Celsius
TSA_Mean	The mean TSA over the entire times period	degrees Celsius
TSA_Frequency	Thermal Stress Anomaly Frequency: number of times over previous 52 weeks that TSA >=1 degree C	TSA per time period
TSA_Frequency_Standard_Deviation	The standard deviation of frequency of thermal stress anomalies over the entire time period	TSA per time period
TSA_FrequencyMax	The maximum TSA_Frequency over the entire time period	TSA per time period
TSA_FrequencyMean	The mean TSA_Frequency over the entire time period	TSA per time period
TSA_DHW	Thermal Stress Anomaly (TSA) Degree Heating Week (DHW): Sum of previous 12 weeks when TSA >=1 degree C	weeks
TSA_DHW_Standard_Deviation	The standard deviation of TSA_DHW over the entire time period	weeks
TSA_DHWMax	The maximum TSA_DHW over the entire time period	weeks
TSA_DHWMean	The mean TSA_DHW over the entire time period	weeks
Date	date of sampling event in format YYYY-MM-DD	unitless
Site_Comments	comments of any issues with the site or additional information	unitless

Sample_Comments	comments of any issue or additional information of sampling event	unitless
Bleaching_Comments	comments of any issue or additional information of bleaching value	unitless