Coral Reef Bleaching Dataset Analysis Report:

**introduction:**

The dataset contains 41361 rows × 62 columns when provided originally. It has a lot of anomalies and with in 62 columns most of them contain the object data type with mixed value.

**Missing values analysis:**

1. The missing values analysis showed different percentage of missing values in different columns with highest missing values in **Site\_Name** column at the ratio of **83**%. This column was simply deleted from the dataset due to almost no available data with in it.

2. Next highest missing values column is **Bleaching\_Level** with missing values percentage of **45**% but will be treated in next steps because 55% of the data is still available so we can give it a try to apply mean, mode or KNN **missing values treatment to evaluate it.**

3. Next highest missing values columns are **Substrate\_Name** and **Percent\_Cover,** with 30% missing values each. These columns can also be treated under the missing value to evaluate.

4. Another column with missing value percentage of **30**% is **Reef\_ID** but this **column will not be treated** because the data in this column is unique Reef\_IDs which are unique to the location with long strings of values with numerical and string values mixed together. This column can be treated under the encoding of dataset by converting the mixed data type into float or integer masking.

Rest of the dataset have no missing value or even less than 1%.

**Eliminating unwanted columns:**

In this dataset some columns were unwanted like Site\_Comments, Sample\_Comments, Bleaching\_Comments, Date, Date\_Day. These are come basic unwanted columns but there can be other unwanted columns too which can be deleted during the data cleaning process.

**Garbage Value:**

The biggest garbage value in this dataset is value ‘nd’ which is never explained or any input about it and its is making a lot of garbage values in the dataset.

It was removed in the dataset and there is no other garbage value in the dataset for the object data type columns.

**Exploratory Data Analysis**

the descriptive analysis shows the numerical columns data analysis in terms of mean, median, mode, count and % metrics according to statistics some of these columns are sensible to analyse such as temperatures frequencies, percentage of bleaching covered, and many others. But some are non-sense metrics such as long and lat averages and counts because they become no location in the world. So we will ignore such non-sense data

for the object data types the count, unique, top and frequency metrics were used which are pretty helpful in creating a clear visual of the object columns. With the conversion of these column types we can see even clear picture of the data types.

In the next step of exploratory data analysis we generated box-plot for the outlier detection which will be treated in next phase

in the same context the data with numerical columns were tested with the bar chart, scatter maps and heat map to evaluate the trends in the dataset. The trends show that most of the data shows the positive relationship. We will discuss some pre-processed graphs here.

1. for the box plot Column ‘Distance\_to\_Shore, Turbidity, Cyclone Frequency, Depth\_m,Percent cover, Percent\_bleaching, ClimSST, temperature kelvin, temperature mean, temperature\_minimum, temperature\_maximum, SSTA, and all other temperature related plots have major outlier which needs to be treated.

2. for the scatter chart reading to relate different metrics with Percent bleaching as y-axis data, we can easily see a lot of positive and some negative relationship between columns, the percent bleaching w-r-t latitude and longitude is strongly connected, but when I comes to the Distance\_to\_shore it has strongly positive relationship, for turbidity it has also strong and tight positive relationship that more the turbidity more the percent bleaching occurs. The data also shows the cyclone frequency relation with the percent bleaching which shows that there are constant and strong relationship between both columns with strongly positive relationship which shows more cyclone frequency the more bleaching is present in the marine life. The percent\_bleaching correlation with the Depth\_m column is highly positive and tightly observed. After the outlier treatment we will evaluate its relationship again. The percent\_bleaching and Percent\_cover strong positive relationship but a huge number of outlier too, which needs the treatment as well. percent\_bleaching and climateSST has a positive relationship that one entity is increased and other increases as well, like if climateSST value is increased the bleaching level is increased as well. The temperature level is also telling the same story. All these tables need the outlier treatment as well.

**Data Cleaning/Outlier Treatment:**

**1. Missing Values Treatment**

there were two ways the missing values of numerical values can be treated one is statistical way of mean, median and mode and another one is KNN imputer method. I have applied both to handle the missing values as numerical values.

**2. Treating Outliers**

the treatment of outliers is done which were identified in the last step of exploratory data analysis. But we target major columns which are developing extreme outliers.

With outlier treatment we can clearly see a difference in results for the Distance to the shore and percent bleaching column with percent bleaching less as the depth is increased and more percent\_bleaching when Distance to the shore is less. **Which shows that less depth and near to the shores the bleaching factor is more active**

turbidity column when given the outlier treatment shows that as the turbidity increases there is less percent\_bleaching occurs, where as the more intense the turbidity will cause more bleaching percent. **This shows that more turbidity causes more percent of bleaching, we need to clean our shores water from huge visible turbidity.**

When cyclone frequency is given outlier treatment its shows more clearer graphs of relationship, linear relationship of more cyclones per terms causes more percent of bleaching in corals.

The relationship between Depth\_m and percent\_bleaching is integral, where more depth is promotion more percent bleaching. **Graph shows that less distance to the shore, the less percent is bleached.**

When percent\_cover is compared with percent\_bleaching, it shows that more percent\_cover on the shores the less percent\_bleaching will occurrent.

When percent\_bleaching is compared with ClimSST after outlier treatment, it shows that with the certain range of the Climate change the percent\_bleaching is occurring at high ratio.  **This graph shows that climate change is contributing huge factor into the bleaching\_percentage**.

The temperature again shows that percent bleaching is impacted by the temperature. The certain range of temperature can increase the percentage of bleaching. **Their graph shows that temperature increase can increase the percent bleaching as well.**

Encoding of Object values

at the end the data types of object columns was converted into int values using the dummy dataset values. These values gives a lot of