**Aim**

Identification of potential fishing zones using machine learning, based on physical and chemical properties of the water.

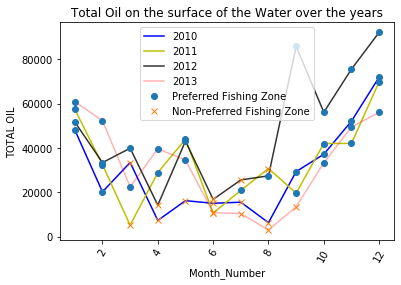
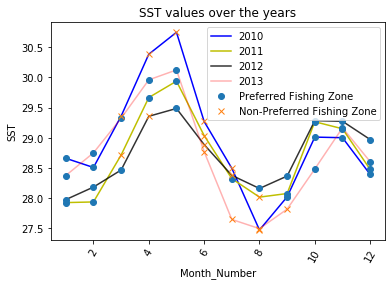
**Population of the dataset**

It was instructed to populate the given dataset via interpolation techniques, we observe that the thin dataset is highly correlative to the seasons of a particular year. To be calculative we got only four year of data across 48 rows.  
We could generate new rows by randomly selecting values of a particular feature say “SST” of a particular “Month” over the four years (similarly for all features) but there could be a contradiction at the point of Labeling its class to either PFZ or NPFZ.

Hence we found it inefficient to populate the dataset

**Data Visualization**

In the given dataset all the seven features SST to TOTALOIL are seasonal features I,e, there values are highly co-related to the present season of the year hence, during each of the four years the rise and drop of the values of any features are statically predictable.



**Use of Up-sampling**

It is observed that the classes (PFZ or NPFZ) in the given dataset are in a ratio of 2:1 which effects the training of the model negatively. Hence we use up-sampling to make the ratio count(NPFZ) / count(PFZ) =1

**Use of ML Algorithms**

As we can observe that the dataset is thin, deep learning algos and reinforcement learning which usually efficient for big datasets will cause overfitting/outliers hence, we may use SVM or decision tree algos like Random Forest.

Hence, we are using Random Forest.

**Results**

The accuracy of the outputs the trained Random Forest on test data is found out to be 85.71428571428572 %

This does not mean that the error cannot be minimized further, we found that when we set the parameter “Random state” to 42 and increase the “estimators” the accuracy of the outputs becomes 100%, this is due to over-fitting not learning hence, it is better to have a learned 85% accuracy in a thin dataset.

were also calculated out “feature\_importances”. We observed that “TOTALOIL” turned out to be the most important feature among all contributing over 60%