Castillo-Sanchez, Sebastian R.

6000 level  Professor Munasinghe

effects OF MARINE coral bleaching on aquaculture productin and key biological areas

**Abstract**

As 80% of the world is covered by water, the resources that the oceans provide comes in large quantities. However, due to exploitation of these key resources and global warming, these resources are diminishing rapidly. Coral reef bleaching is one of many resulting effects and it could have a long-lasting impact to the oceans. Different types of marine life rely on these coral reefs for food, general ecosystems, etc. . In turn, many people around the world rely on marine life and coral reefs for food, jobs, etc. Therefore, it should be important to analyze the effects of coral reef bleaching on areas such as aquaculture production and biodiversity. An article from *Nature Communications* published in March of 2019, states that “Thermal-stress events associated with climate change cause coral bleaching and mortality that threatens coral reefs globally.” (Sully, Burkepile, Donovan, Hodgson & van Woesik). The problem that this project will dive deeper into is if marine coral bleaching present in different countries around the world affect the country’s aquaculture production as well as their respective Key Biological Areas. With climate change as well as global warming increasing largely every year, we can hypothesize that marine coral bleaching is affecting marine life. As the UN has specified “Life Under Water” as one of their Sustainability Goals for the next few years, there is starting to be more insight into what is happening around the world. Key Biological Areas have been described as “sites of global importance to the planet’s overall health and the persistence of biodiversity” (KBA.org). In addition, aquaculture production refers to “output from aquaculture activities, which are designated for final harvest for consumption.”(Ritchie, Roser). Furthermore, aquaculture harvests fish, mollusks, crustaceans, etc. which are sources of food for millions around the world. Therefore, the motivation is to understand the underlying threats that marine coral bleaching has on the biodiversity of marine life and how it could affect society in terms of economy and nourishment. This report will give an overview of the problem being looked at and its approaches. A brief introduction about the data sources used as well as how the datasets were used alongside each other. Next, explanatory data analysis (EDA) will be shown to explain the data in a more visual way with plots and graphs. Then, the process of modeling, optimization and tuning will be shown to further analyze. Lastly, there will be conclusions and discussion about the results and further work.

**Data Description and Exploratory Data Analysis**

Three different datasets were used for this project to cover all aspects that will be used for analysis as well as modeling. The first dataset is from BCO-DMO (Biological & Chemical Oceanography Data Management Office) and is called “Global Bleaching and Environmental Data” The investigators who collected this data are Robert van Woesik from FIT and Deron Burkepile from USCB. They are also authors of the article from Nature Communications that was cited in the Abstract This dataset consists of different climate and environmental data collected from across the world to measure coral reef “bright spots” throughout several years. This consists of attributes such as “Windspeed”, “SST”, “Climate”, “Temperature”, etc. This data also contains amounts of coral bleaching in different regions, oceans, countries, etc. All these environmental attributes could also influence bleaching so analysis will be done measuring the correlation between them. This dataset contained 9665 observations (rows) with 48 variables (columns).

The second dataset comes from Our World in Data which is a part of the Global Change Data Lab based in the United Kingdom and a joint venture with the University of Oxford as well as Oxford Martin School. The specific data comes from an article titled “Seafood Production” by Hannah Ritchie and Max Roser. The basis of this article is to measure the production of seafood globally by different methods. In addition, it analyzes what is being overexploited, and what method is producing the most seafood around the world. It also provides specifics in production with those methods with one of them being from Aquaculture. The data they use to analyze comes from the UN Food and Agriculture (FAO) general database as well as their FishStat database. For this project, the specific data being used is the “Aquaculture production” dataset which measures Aquaculture production in different countries by metric tons. Below is a visualization from the article measuring global Aquaculture production by different regions:

Chart, line chart

Description automatically generated

*Figure 1: Aquaculture production by Region*

This data will be very useful in measuring the consumption in different countries and see if coral bleaching is affecting any of the production around the world. This dataset contained 10880 observations with 4 different variables.

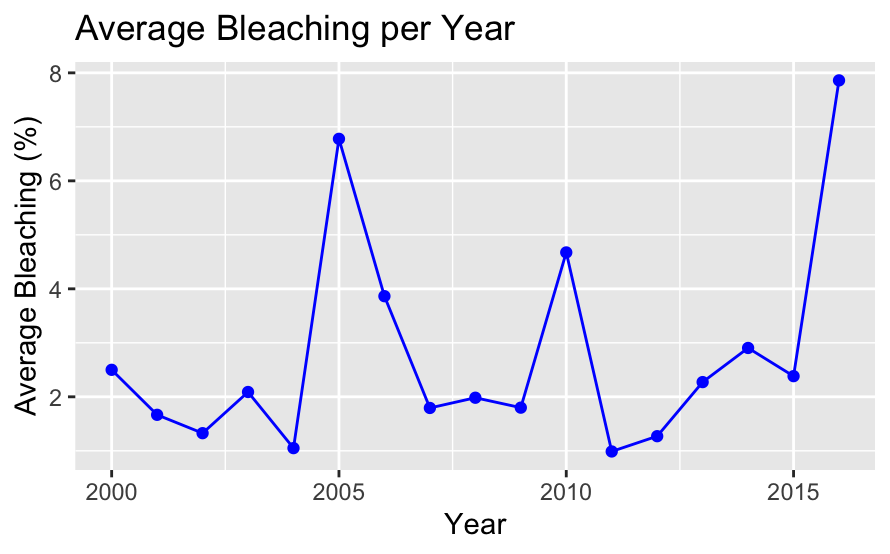
The third and final dataset being used is from the UN Sustainable Development Goals website. As said before, one of the goals (#14) is “Life Under Water” which is a goal to improve upon life under water, improve biodiversity and keeping the marine environment clean. The specific indicator being used for this project is Indicator 14.5.1: Coverage of protected areas in relation to marine areas. The specific name for the dataset is “Average proportion of Marine Key Biodiversity Areas (KBAs) covered by protected areas (percent)” (UN SDG) As also explained before, KBAs are very important marine life areas that provide a lot of biodiversity and have an impact to the surrounding ecosystems. The UN’s goal is to conserve these areas little by little in the next few years. The attributes that this dataset contains are countries and percentage of KBA coverage from 2000 until 2019. The dataset contained 169 observations with 103 variables.

Overall, these three datasets will provide more than enough information for further analysis as well as modeling. Even though they come from different sources, they have specific information that will be able to be analyzed in different ways and provide insight into what is happening all around the world.

**Analysis**

After compiling all the different datasets and reviewing what they contained, it was time to perform some analysis. However, before analysis there had to be some data cleaning as well as some merging done for ease and efficiency. For cleaning, a lot of data needed to be converted to numerical data as the numerical data in some of the attributes was specified as a character variable. The coral bleaching dataset had a date format for their data, so I converted the date column into a date type as well as extracting the year into a new column for merging later. Since all the data in this dataset start from 2000, the merging of all three datasets will rely on the last two decades to have more unified, clean data. Since the numeric data was what was important, I decided to remove character columns as well as positional data as it was not going to be needed during modeling. However, I kept the character column that was similar across all three datasets which was the one that contained the country names. This column alongside the newly created “Year” column were the indexes that would help during merging. For data synthesis (merging) I incorporated the “dplyr” package in R. For the coral bleaching data, I grouped the data by country name and year while merging the rest of the attributes by the mean. For further context, since various countries had different values for the same year, they would be grouped and then the average was found for the environmental data. The first merge consisted between the coral bleaching and the aquaculture production data using the “inner\_join” function by the country name and year indexes. Lastly, it was performed again with the UN KBA dataset to create a fully merged dataset. The UN KBA dataset format had different columns for each of the KBA percentages by year, so it had to be flipped into rows before merging. Null values were handling during the cleaning of each of the datasets respectively so the merged data did not contain null values which would help for modeling/analysis.

Since merging was done, analysis could be done to explore the statistics of the data. The first two analyses performed was to see the trend of average coral bleaching per year as well as the average KBA per year. The visualizations are below:

**Chart, line chart

Description automatically generated**

*Figure 2: Average Bleaching per Year Figure 3: Average KBA per Year*

Figure 2 shows that there is a lot of rises and drops when it comes to the percentage of average coral reef bleaching throughout the years. There are huge rises in 2005, 2010 and 2016. This could be due to rising temperatures across the world or temperature of the ocean surface. However, the rise in 2016 is the largest and most recent rise which should bring some concern and worry. Figure 3 shows that there is a positive result in the average KBA percentage per year. From 2001, the KBA percentage increased by almost 30% and has slowly increased ever since. We can conclude from this visualization that there is more focus on increasing the percentage of KBA all around the world and the UN goal is obtainable. Since there were a lot of other numerical attributes in the data, it was important to also analyze the distributions of several key indicators. Below are four different distributions of these indicators:

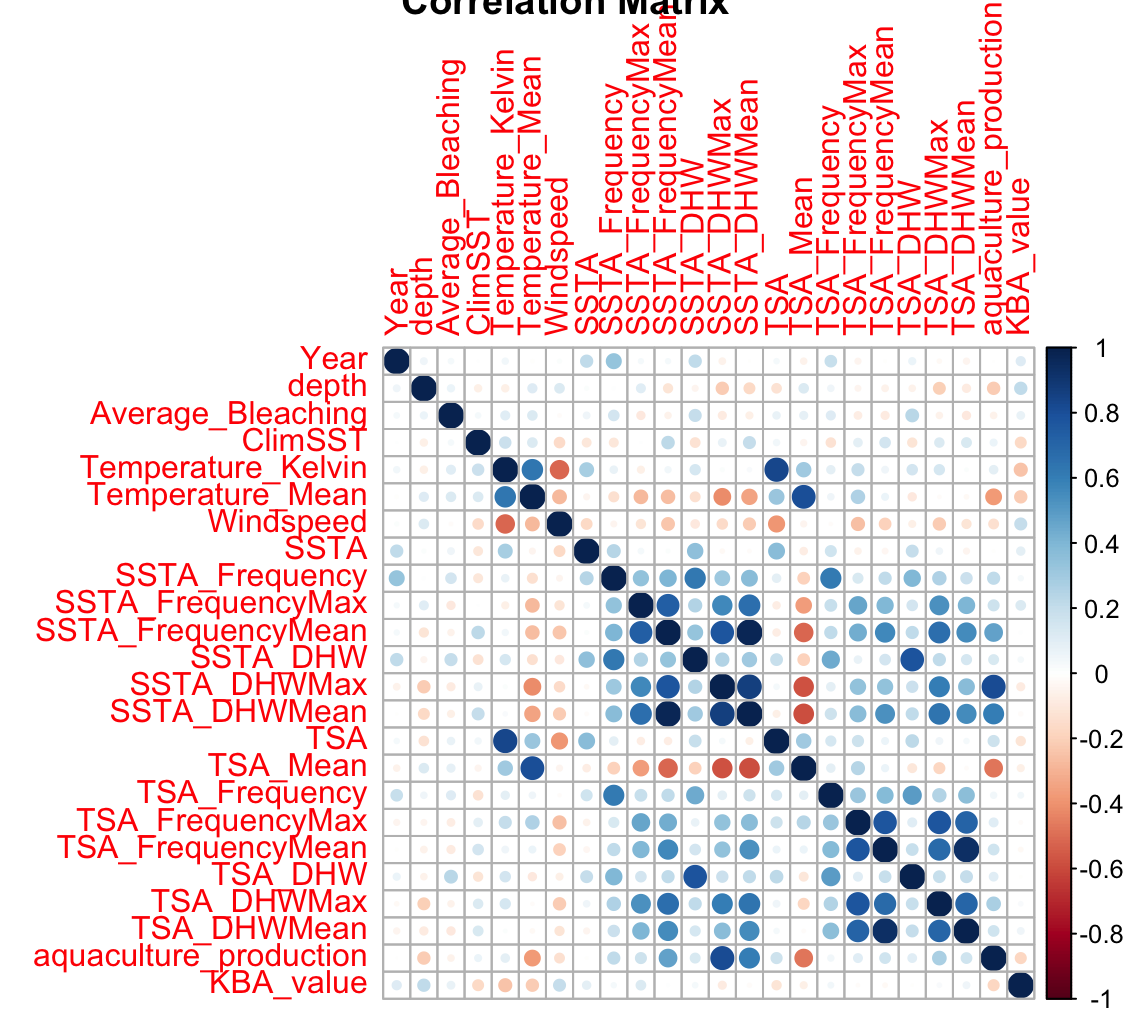
**Chart, histogram

Description automatically generated**

**Figure 4: Distributions of important attributes**

The first histogram (red) is a distribution of the Temperature\_Mean column that measures the average temperature in degrees Kelvin. The distribution is more rightly skewed which means more of the temperatures are high (hotter). Most of the data is in the 300-302 range which equates to around 80 degrees Fahrenheit. The second distribution (green) measures the windspeed in meters per hour. This one is more of a normal distribution. The blue colored distribution measures the distribution of SSTA which is described as the Sea Surface Temperature Anomaly frequency. This attribute measures the weekly Sea Surface Temperature (SST) subtracted by the weekly climatological SST. This is a very normal distribution. Lastly the distribution (black) measures the distribution of the depth attribute which is the distance from surface to study site (in meters) according to the data description. This distribution is a little more left skewed which means more of the observations are more shallow depths.

Another analysis that was very important in helping to understand the data was to measure the correlation between the variables. Plotting a correlation matrix will allow me to visualize the relationship between variables and see if any of the variables impact others. Figure 5 below shows the completed correlation matrix using the “corrplot” R package:

****

*Figure 5: Correlation Matrix*

The correlation matrix provides different levels of correlation amongst all the numeric variables. From the plot, some interesting insights that come out is that neither of the Average\_Bleaching or the KBA\_values columns have very high or very low correlations amongst them. We can conclude that none of the environmental variables have an impact on coral bleaching or KBA percentage which is interesting as a lot of them could impact. The aquaculture\_production variable has more prominent correlations with a high correlation with the SSTA\_DHWMax which is Sea Surface Temperature Degree Heating Weeks maximum value over a time period. There is also a low correlated variable which is the TSA\_Mean which is the mean value of Thermal Stress Anomaly over a time period.

Lastly, determining sources of error and uncertainty in the data is very important in understanding the data fully. After the data was fully merged, the data size condensed to about 330 observations with 27 variables. This was due to it only being merged where there were similar countries and similar years. The data in the coral bleaching did not have set years for every country/region while the other two datasets did. There could be error due to not having enough data in our dataset, however after looking at the analysis and EDA, the data shows no signs of difference or loss due to low amounts of data. Also, from looking at the correlation matrix, it seems that some of the important variables such as Average\_Bleaching or KBA\_Value do not have much correlation with other variables therefore it will be important to see if they are significant using modeling.

**Model Development and Application of Model(s)**

After statistical analysis was performed to better visualize and draw insights from the data, it was time to move onto the development of different machine learning models and the application of these models. Since all the variables in the full dataset that we wanted to focus on were mostly numeric as well as continuous, regression analysis was the best method to create models for this data. To find the best results for the data, three different regression models were developed. Then after performing regression and analyzing the results, a clustering analysis was performed using the data.

For the following regression models, a train-test split was created for the numeric only data. The split created using the sample function had a 70-30 percentage split for the train and test data. This was done to train the models on the training data while testing it after with the test data while creating predictions.

The first regression model performed was a linear regression model. Linear regression measures the relationship between two different variables, x (independent variables) and y (dependent variables) . For this dataset, three different linear regression fits were created that analyzed different relationships. The three different specifications of the models are listed below:

* Fit 1: y = Average\_Bleaching , x = Rest of numeric variables (train set)
* Fit 2: y = aquaculture\_production , x = Rest of numeric variables (train set)
* Fit 3: y = KBA\_value , x = Rest of numeric variables (train set)

Fit 1 first measures the significance of the numeric variables of the data on the dependent variable which is Average\_Bleaching. As explained before, various conditions such as climate, temperature, etc. could have a large impact on coral bleaching. Fit 2 measures the significance of the numeric variables on the aquaculture\_production data. This will see if any environmental data or more importantly the coral bleaching data will have significance. Lastly, Fit 3 measures the significance of the numeric variables on t KBA\_value percentage

The results for the Linear Regression fits are below in the table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Fit** | **Multiple R-Squared** | **Adjusted R-Squared** | **P-Value** | **Significant variables (noted by Significance codes)** |
| **Fit 1** | 0.2123 | 0.1235 | 0.0006 | TSA\_DHW \*\*\* |
| **Fit 2** | 0.8819 | 0.8686 | < 2.2e-16 | TSA\_DHWMax \*\*\*  TSA\_FrequencyMean .  SSTA\_DHWMax \*\*\*  SSTA\_FrequencyMax \*\*\*  Year \*  (Intercept) \* |
| **Fit 3** | 0.3235 | 0.2473 | 7.994e-09 | TSA\_DHWMean .  TSA\_DHWMax \*\*  TSA\_FrequencyMean \*\*\*  TSA\_FrequencyMax .  SSTA\_DHWMax \*\*  ClimSST \*  depth \*\*\*  (Intercept) \* |

*From R output: Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

From the results from all the fits, there can be a lot of insights generated that could help with the problem this project is looking at. Fit 1 shows that with its two different R-Squared values, it does not explain the data well with an Adjusted R-Squared score of only around 12%. This means that the regression model only explains less than 12% of the data which is not very good. With the significance, you can see that TSA\_DHW (Thermal Stress Anomaly Degree Heating Weeks) is the most significant variable that affects the Average\_Bleaching variable. The other variables that are listed as significant have low to no significance. This is an interesting insight as more variables could’ve had a larger significance, but it shows that Thermal Stress could affect bleaching amounts. Fit 2 is opposite to Fit 1 as it has much higher R-Squared values with an Adjusted R-Squared of around 86%. This regression fits the data much better and explains the variance more. This fit had more significant variables with the most coming from TSA and SSTA variables. We can conclude that aquaculture production is impacted by attributes such as Thermal Stress Anomalies and Sea Surface Temperature Anomalies. This could mean that warmer water affects the aquaculture production in various areas where the climate is warmer or colder. The third and final Linear Regression fit, Fit 3, had a similar low R-Squared scores with an Adjusted R-Squared score of around 25% which is higher than Fit 1 but still low explainability. The significance variables are like Fit 2 with the most significant variables are TSA or SSTA related however a difference is that Fit 3 has the depth variable as highly significant here. This could mean that the depth of the site can affect the percent coverage of KBA as deeper areas could have more biodiversity. For a deeper insight, two more linear regression fits were performed on Average\_Bleaching on both aquaculture\_production and KBA\_value by themselves that resulted in no significance between them (see code). The Average\_Bleaching variable estimate was negative on the KBA\_value regression which could mean increase in KBA\_value means decrease in Average\_Bleaching. These small fits shows that there are no large significant effects for coral bleaching on aquaculture production and KBA coverage. However, overall, the Linear Regression fits gave more insight into the variables and how they impact each other at a basic level.

The second regression model that was performed was a Decision Tree regression model. Decision Tree models use observations and the target variable for decisions. The branches in this case are the observations while the leaves are the target variable. For this project, I decided to use the “caret” R package which makes it more efficient to perform machine learning models. The method for decision trees incorporated the “rpart” package and the train function is what controlled the modeling. For the best results, cross-validation is performed in the modeling process using the trainControl function. This cross-validation consisted of a 10-fold cross-validation that repeated 3 times. From this, two different regression decision trees were performed like what was performed in the Linear Regression. The package “rpart.plot” was used to visualize the decision trees. Below are the visual results of all three decision trees:

**Timeline

Description automatically generatedTimeline

Description automatically generated**

*Figure 6: Decision Tree Models*

The first plot in Figure 6 (left) is the Decision Tree Model of the KBA\_value as the y variable while the rest of the numeric variables are the x variables. You can see that in this model, the aquaculture\_production variable plays a huge part in splitting the tree as the first decision. If the value of the column is less than 22e+3 than 50% of the data is in that leaf. If not, it splits again based on Windspeed where if it is less or greater than 4.8 meters per hour. In this split, it is basically an even split with 23% “yes” and 24% “no”. This model is very interesting as it shows that the amount of aquaculture production impacts the KBA\_value as well as Windspeed level. This shows that it is important to perform several different regression models to see different but insightful results. The second Decision Tree model (right) is the aquaculture\_production as the y variable and the x as the rest of the numeric variables. This tree model has SSTA\_DHWMax as the first split decision variable. This is an interesting result as this variable was one with a lot of significance in the Linear Regression fit. The second split on the “yes” side is on TSA\_Mean being less or more than -1.1. Most of the aquaculture production data (84%) is resulting on the “yes” side for this variable meaning that SSTA\_DHWMax and TSA\_Mean has an impact on the aquaculture\_production variable. There is a third split on the Year column being less or greater than 2012 however there is a low amount of data in this split. As you can see, Average\_Bleaching is nowhere to be found in these models which again shows that there isn’t a large impact between what we are trying to compare. Overall, these two models were important in gaining more information about what variables impact these different variables.

The third and last type of regression model performed was a SVM (Support Vector Machine) Regression model. Support Vector Machine is a machine learning model (supervised learning) that focuses on class separation. It is normally used for Classification models however for this project, it is being used for Regression modeling. Like the Decision Tree models, the caret R package is used for modeling. The method for SVM is using “svmLinear2” from the e1071 R library. The same cross-validation method from trainControl is being used with 3 times repeated 10-fold method. The two models created for SVM Regression were one for the KBA\_value and another one for the aquaculture\_production variable. A Linear Kernel was used for both models for simplicity. Results for the best performing SVM Regression for each model after cross-validation are shown below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Cost** | **RMSE** | **Rsquared** | **MAE** |
| **KBA\_value ~ .** | 0.5 | 19.87980 | 0.259 | 14.434 |
| **Aquaculture\_production ~ .** | 1 | 3355048 | 0.687 | 1772336 |

The results for these models aren’t as visual as the Decision Tree or not too definitive like the Linear Regressions but they are still important to analyze. The comparison between Rsquared scores is somewhat of a large difference as the second model is close to 69% while the first model is only about 26%. The parameter Cost is different which means that the cross-validation These models can’t show significance however it can be very useful when creating predictions.

For all these regression models, predictions could be created using the training fits and it being tested on the testing set. For the train sets, each model was trained using 224 observations and the testing/predictions was done on 98 observations in the testing set. The predictions were done using the respective models as well as the predict function in R. Since this was a regression problem, the predictions are not done by accuracy but by plotting the actual values on the predicted values. The results of all the predictions are explained in the plots below:

Chart, scatter chart

Description automatically generated

*Figure 7: Linear Regression Prediction Results*

The results above from Figure 7 show the different qplots for the results of the predictions of the different fits. All the plots are the actual values plotted against the predicted values (test vs predicted). The Fit 1 plot shows the Average\_Bleaching values colored by year. The plot shows most of the values together between the 0-10 values for the test values while between -5 and 10 for the predicted values. This plot does not have a linear relationship which could mean there is not enough data in the testing set for predictions or the predictions don’t show a linear relationship. In addition, the color of the points varies therefore there isn’t a specific year(s) that are more frequent than others. The Fit 2 plot has more of a linear relationship for the predictions for the KBA\_value but it is a lot more scattered than Fit 1. When it comes to color, it seems that most of the values in the plot are darker which correlates to lower levels of Bleaching. Fit 3 is a very condensed plot when it comes to points. This could be due to the range of the plot where several of the points are much larger than the rest. The color of all the condensed points seems to be darker which correlates to lower levels of bleaching.

**Chart

Description automatically generated**

*Figure 8: Decision Tree Prediction Results*

Figure 8 shows the visual results of the Decision Tree model predictions. The two fits were the two explained earlier. These two plots are very non-linear and seem to have a line of values for specific values. There is not much to conclude as the plots do not show any trends or linearity. However, for the color of the points, most are lower levels of bleaching which is like the Linear Regression predictions and an interesting insight

**Chart, scatter chart

Description automatically generated**

*Figure 9: SVM Regression Prediction Results*

The last plots for the regression models shown in Figure 9 show the results for the SVM regression model fits. The first fit (Fit 1) shows a more linear relationship between the test values and the predicted values for the KBA\_Value. There is some scatter to the data but there is somewhat of an upwards trend. The second plot (Fit 2) is like the same fit from the Linear Regression model. Since there are a few large values in there the rest of the values are condensed together in the lower left corner but no linearity to be seen. As for color, like the other models, most of the points show low amounts of bleaching.

The last modeling performed on the data is a Clustering analysis. More specifically a K-Means Clustering was done using the data. For a background, K-Means clustering is a simple Clustering method that focuses on finding relationships in the data with no training models. The K in K-means is the number of groups that the data is split into. Clustering is a popular unsupervised learning method. Before performing this type of modeling analysis, there a split done in the Average\_Bleaching column to split it into classes to visualize against the KBA\_value and aquaculture\_production variable. Figure 10 below shows this visualization:

**Chart, scatter chart

Description automatically generated**

*Figure 10: Average Bleaching classes by KBA\_value and aquaculture\_production*

The classes are split into three different types being “low”, “medium”, and “high” depending on the value of Average\_Bleaching which was 0-5, 5-20, 20-80 respectively. This visualization is very helpful in visualizing the how each of the variables affect each other. There are a few points that have a low KBA\_value coverage percentage but high amounts of aquaculture production with low amounts of bleaching. There are also points with very low amounts of aquaculture production with high levels of coverage of KBA percentage that have low amounts of bleaching. KBA percentages that are high have very low amounts of aquaculture production which is a very interesting insight. This insight was also shown during the first Decision Tree model fit. Many of the points in this plot are either low or medium levels of bleaching with only a very little amount of high bleaching. To find the optimal number of clusters, a plot finding the within sum of squares was created. The plot showed that 3 clusters could be the most optimal number of clusters. The K-Means Clustering was done using the kmeans R function using normalized (scaled) data of the numeric data and 3 clusters. The data was then aggregated using the mean function and aggregated by the different clusters that were created. Below is the final plot:

**Chart, diagram, scatter chart

Description automatically generated**

*Figure 11: Visual results of K-Means Clustering*

Figure 11 shows the three different K-Means clusters and all the points that are put into each of the clusters. The components from the plot explain 42% of the point variability. For the most part, all the points are in their respective clusters. Clusters 2 and 3 seem to have some overlap for some of the points however, the clusters do seem to have separated the data well and there are no problems with using 3 clusters.

**Conclusions and Discussion**

Overall, this project gave insight to different environmental and marine data that has an impact to life under water as well as life above it. Coral reef bleaching is a real situation happening in the world right now and there should be something done about it. Climate change as well as global warming are huge factors that contribute to bleaching and will continue to have an impact if nothing is done about it. This project was to see if there were any effects of marine coral bleaching on aquaculture production as well as coverage percentage of Key Biodiversity Areas (KBA). From looking at the results from analysis and several different types of models, the conclusion is that there is no clear impact of the average amount of coral bleaching in the countries that were in the dataset on these two areas. However, the results from analysis and modeling showed that differ One of the problems with the data was after cleaning/merging, there was a significant drop in total observations (rows) to only around 330 rows. There were some countries that were left out due to them not being in all three datasets as well as there were some years that had data for some countries and years that did not. These could have contributed to the opposition of the original hypothesis that there would be a significant impact from coral bleaching. In the future, having a larger dataset that contained most if not all the countries surrounded by water for equal number of years for each will allow for better analysis, modeling and predictions.

**Code**

GitHub: <https://github.com/scastillosanchez/Final-Project-DataAnalytics>

Caret package documentation: <https://topepo.github.io/caret/model-training-and-tuning.html>

**References**

Sully, S., Burkepile, D.E., Donovan, M.K. et al. A global analysis of coral bleaching over the past two decades. Nat Commun 10, 1264 (2019). <https://doi.org/10.1038/s41467-019-09238-2>

Hannah Ritchie and Max Roser (2019) - "Seafood Production". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/seafood-production' [Online Resource]

van Woesik, R., Burkepile, D. (2019) Bleaching and environmental data for global coral reef sites from 1998-2017. Biological and Chemical Oceanography Data Management Office (BCO-DMO). (Version 1) Version Date 2019-07-18 [if applicable, indicate subset used]. doi:10.1575/1912/bco-dmo.773466.1 [access date]

<https://unstats-undesa.opendata.arcgis.com/datasets/indicator-14-5-1-average-proportion-of-marine-key-biodiversity-areas-kbas-covered-by-protected-areas-percent-5/data?geometry=136.385%2C-32.368%2C-133.615%2C66.033&page=6>

<http://www.keybiodiversityareas.org/about-kbas>