**Capstone Project 1**

**Has Climate Change had an Impact on Commercial Fishing**

**Kevin Bryson**

**Springboard**

**May 11th, 2018**

1. **Introduction**

Imagine that you are about to make a fantastic dinner consisting of your favorite fish. You drive down to the local fish market or grocery outlet and to your surprise and disappointment, the fish you wanted is no longer available. After several inquiries to the person behind the counter you learn that particular fish, no longer exists or has been classified as endangered and can no longer be sold.

The Earth’s oceans have been undergoing hydrochemical changes for quite some time. The ocean is absorbing Carbon Dioxide at an increasing rate. This increase in Carbon Dioxide can have a lasting effect on all aquatic life in the ocean including fish. The problem is to what effect is this increase having on fish populations.

Governments and commercial fisherman need to understand how these changes should impact not only policy but also the livelihood of every one of those fishermen.

1. **Data Acquisition and Processing**

I searched the internet for my datasets and got them from [www.noaa.gov](http://www.noaa.gov).

The data covers the years from 2000 to 2017.

The datasets were then transferred to a jupyter notebook and then loaded using pandas and the read\_csv function. After checking the data for any NaN values, the strategy chosen to deal with NaN values was to use the mean for that subset. After that was taken care of the data was then grouped by year and species name and then reindexed to allow each column to be used. The data set containing the Carbon Dioxide values need further cleaning so that the units were only in the column name and not used in each value. After splitting the datasets into years from 2000 to 2017 I then used the merge function to put the three data sets together, calling them m2, using the year column for commonality.

The matrix m2, has 126 rows representing each of seven fish species for our given years and 6 columns that include:

* CPUE- Abundance of the Fish Species
* RPN- Total number of all fish divided by the total number of stations
* RPW- Total fish biomass for each species
* Total Number Caught
* Mean Weight (kg)
* CO2 ppm- Amount of Carbon Dioxide in the ocean

1. **Exploratory Data Analysis**

As mentioned above, there are 6 columns in our dataset and each row consists of one of seven species for a given year:

* Giant Grenadier
* Pacific Halibut
* Pacific Cod
* Rougheye Rockfish
* Sablefish
* Shortraker Rockfish
* Short Spine Thorny Head

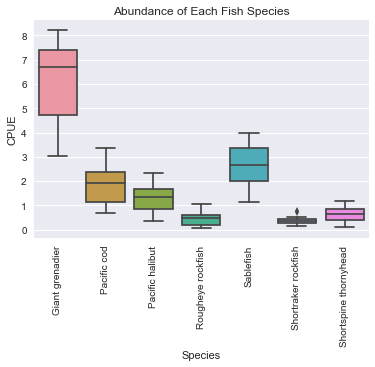
Our goal is to support or reject our Null Hypothesis that:

* There is no significant statistical difference between the amount of carbon dioxide in the ocean and the number of fish and the mean weight for each species.

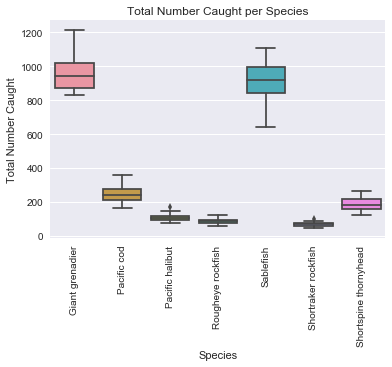
If we see that there is a statistical significance, then we could reject the Null Hypothesis and support our alternate hypothesis:

* There is a statistical significance between the amount of carbon dioxide in the ocean and the number of fish and the mean weight for each species.
  1. **Examining the Numbers of Fish**

The CPUE value is the abundance of each fish species. According to our data, each species has a differing amount of abundance with the Giant Grenadier having the largest and both rockfish species being the least.

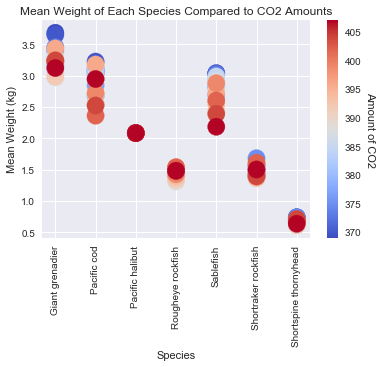


This can be seen using the following boxplot. One of the interesting things about this plot is that the higher the abundance of the fish species, the more variation we have in the actual value. We could predict that the more abundant the fish species is, the more fish that would be caught. We can investigate this with a similar illustration.



It looks like the number of Sablefish that are caught is very similar to the number of Giant Grenadier even though the Grenadier has a higher abundance. Sablefish may have less predators or their populations could be centered around larger fishing zones.

Looking at the fish total data and comparing that to carbon dioxide these seems to be some interesting trends.

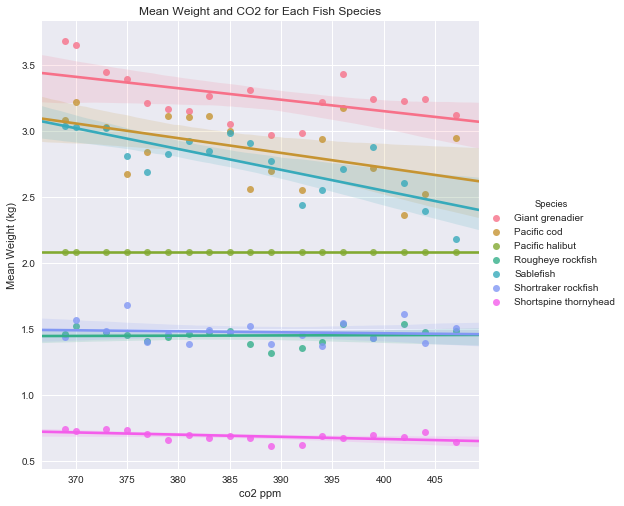


According to this example, Mean Weight seems to be higher when the CO2 levels are lower(blue). There are a few exceptions, Pacific Halibut seems to have little variation and Rough Eye Rockfish seems to vary little depending on the amounts of carbon dioxide.

Examining our data further, we can check for correlation and calculated p-values. The data is not normally distributed so another correlation method other than Pearsonr correlation coefficient will need to be used. Spearmanr correlation coefficient, which is rank order correlation, uses the direction and strength of two ranked variables. For us to be able to reject our Null Hypothesis, we would need to see strong correlation with our data and p-values under 0.05.

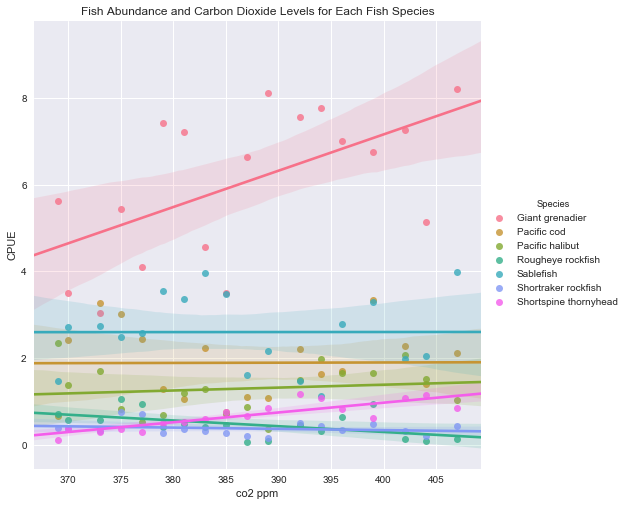
One thing to keep in mind when you are checking for correlation, is not to forget to graph your data. Graphing can allow you to see what you have calculated depending on what correlation coefficient you have used. You should never rely on just the calculated value because the graph might highlight some interesting trends that the numbers just don’t. Evaluate your p-values and compare them to your Null hypothesis.

Looking at mean weight vs. carbon dioxide levels we have a spearman value of -0.077 and a p-value of .39. This tells us that there is no correlation between these to variable and that part of our Null Hypothesis cannot be rejected.



Visually, the graph shows that most of our fish species have a negative slope. The Mean Weight decreases as the carbon dioxide level increases for almost every fish species most noticeable the Sablefish and the Giant Grenadier.

Examining the abundance value of the fish, CPUE, in relation to carbon dioxide amount, we have a spearman correlation value of 0.042 and a p-value of 0.64.



This graph shows some interesting trends as some of the CPUE slopes are increasing or positive such as the Giant Grenadier and the Pacific Halibut. Others are flat and have no correlation and some are negative or decreasing.

Finally, the total number caught of each fish species vs. carbon dioxide levels has a spearman correlation value of 0.00 and a p-value of .99. There is no correlation at all between these two variables and a very high p-value indicates that this variable does not help us reject the Null Hypothesis.

The other two variables, RPN and RPW, have spearman correlation values of 0.04 and -0.02 respectively. RPN has a p- value of 0.63 and RPW has a p-value of 0.82.

**4.Model**

Understanding which fish species can be determined from an unknown data source would useful to expand our dataset. We can use a logistic regression to accomplish this. Before we begin we have should preprocess our dataset and find the features (columns) that have the greatest influence on our model. We can then split our data up into a training set and a testing set to determine the accuracy of the model. Then we can fit those two into our model to make predictions.

LabelEncoder can be used to put our columns into easy to understand data accounting for any subsets in our data. Next, we need to separate our columns into features and a target. Then we can standardize the data between 0 and 1 to make the model run smoother.

To Determine which features are the most important, we can use RFE to accomplish this. We can choose the number of features we want inside RFE(). Four features were determined to be the most influential, CPUE, Total Number Caught, Mean Weight (kg), and co2 ppm. Since logistical regression is being used we can find out which value of the hypervariable C is the best using GridSearchCV(). Using the features from the RFE, and cross validation of 3 folds, the best value for C was determined to be 10.0.

To set up logistical regression, we need to import it from sklearn.linear\_model. The target variable does not only have 2 possible outcomes, there are seven so when logistical regression is passed as logreg, inside the parenthesis we need to specify the multi\_class argument that multinomial is needed and the solver argument is newton-cg.

First attempt at creating the model, the accuracy score was .54 or 54% accurate. None of the data was standardized and the optimal hyperparameter value was not determined. After the data was encoded and then standardized and ran through GridSearchCV, the accuracy rate improved significantly to .94 or 94%. We used a training value of 30 percent of our data and tested the model over 70% of the data to determine its accuracy.

**Trends and Evaluation**

Using the visuals, and the Spearmanr correlation coefficient, some of the data did show positive correlation and fell below our p-value threshold of 0.05. When abundance, CPUE, and the total number of fish caught per year was evaluated, we saw a positive correlation with a Spearman correlation value of 0.83 and an extremely low p-value of 1.4e-32. Meaning that between those two variables there is in fact some statistical significance.

Other data that would be helpful to get a better understanding of how climate change is influencing commercial fishing would be temperature, pH, price of fish accounting for inflation and amount or the number of fisherman. The number of fisherman number may lead to higher number caught and influence the results. Temperature and pH would be important because as the earth gets warmer we would expect to see an increase in water temperature and the increase in carbon dioxide would in fact have an impact on the pH.

**Conclusion**

According to our data there seems to be little or no correlation between number of fish caught, mean weight and abundance of fish compared to an increase in carbon dioxide levels. One of the variables, mean weight did show a negative correlation but the p-value was above the threshold of 0.05. Increasing our sample data would give us potentially better results. Also not limiting our region to Alaska might also show better results. As of now it is impossible to reject our Null Hypothesis as our data analysis was inconclusive. It would seem that as the climate in the Gulf of Alaska has changed, there seems to be no reason that fishing would be impacted. Further study is necessary.