How I organized the datasets (following this github post):

<https://github.com/b7tran/calcofi_hackathon>

Quick summary:

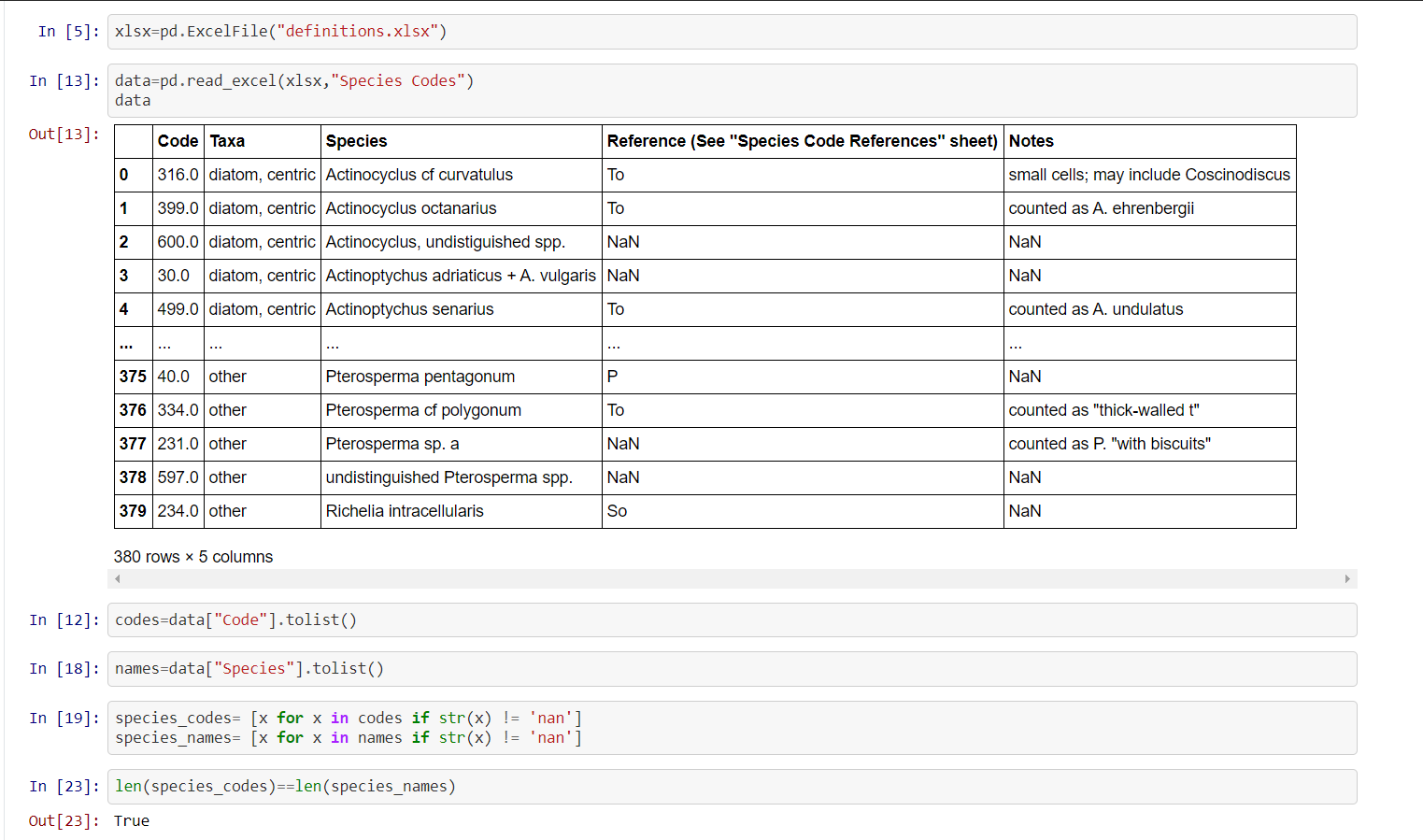
The excel dataset I was working on was already cleaned and formatted by Charlene, who did things such as pivot the tables and clean up the dataset. I manually combined the datasets through excel and then worked on changing the column names for the species from there.   
  
In my first try, I was able to format a dictionary through python, which takes input and returns whatever value is associated with it. In this case, I set up species codes and names in the dictionary so that writing in a code would output the names of the species. Using this, I then worked on using this to change the column names in the datasets, so that we would have the actual names rather than just a bunch of numbers as the column titles.

In my second attempt, I noticed that some of the sheets had very different column names and couldn’t be combined nicely, and worked on removing any outliers, such as column names with only a species number and no associated name, or column names not found in all the excel sheets.

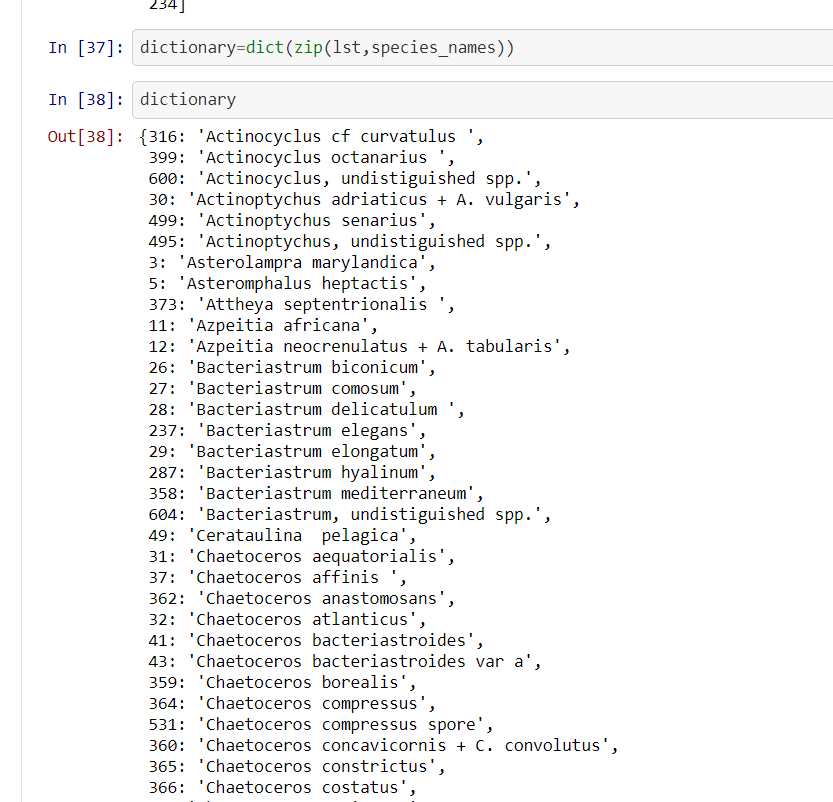
After everything was formatted correctly, and each year had the exact same column headers, I then manually combined all the datasets into one big sheet, so that there could be a complete guide of phytoplankton catches from 1996 to 2018.

Setting\_up\_datasets\_diff\_years\_first\_try.ipynb:

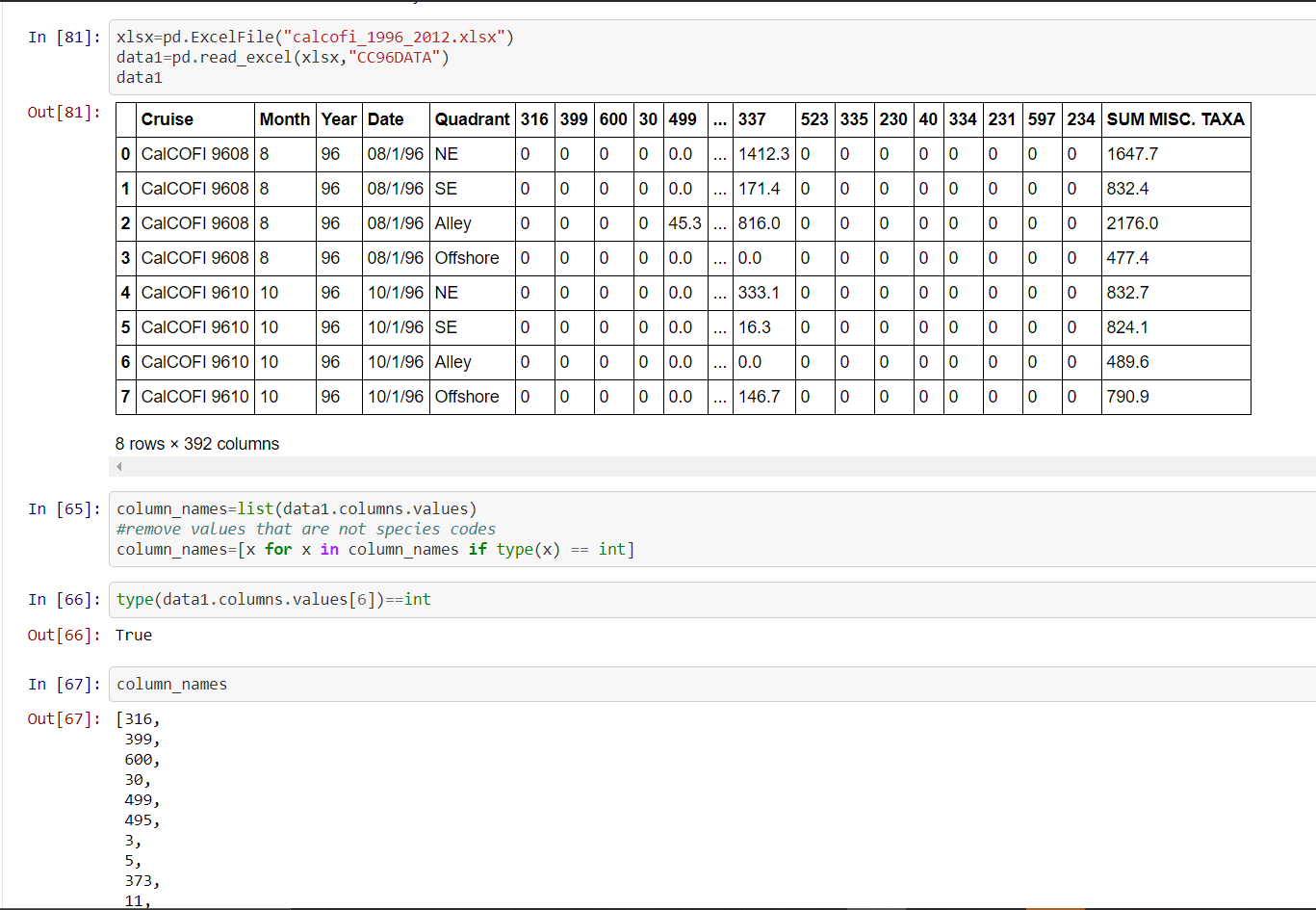
First, I needed to set up a dictionary, which holds keys/values that could associate the species code with their actual names. These values were taken from the “definition” excel sheet, and using Python “pandas”, I got the following columns and set them to equal length lists



Combined the two lists by zipping them into a dictionary. Now, any time a number was inputed into the dictionary, it would return the actual species’ names



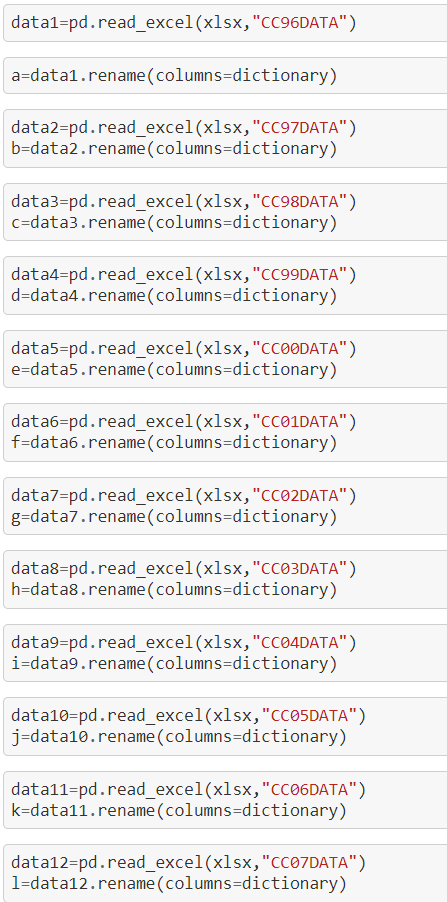
Extracted column names from a single sheet from the workbook. Made sure to set parameters to only get integers, since I just wanted to change the species codes



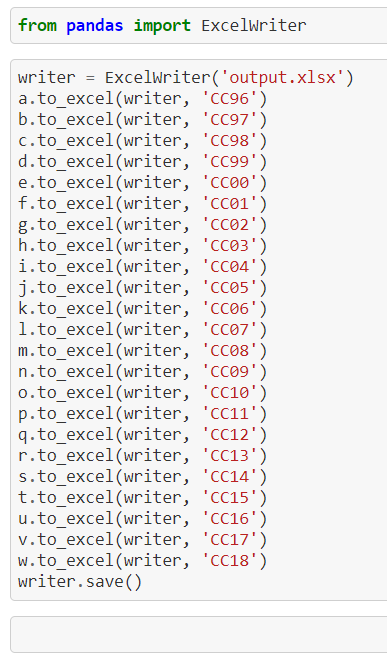
Using another module named “openpyxl”, which allows me to do things like read through excel sheets, I made a list of the sheet names within the excel file.



Not to familiar with how to loop through this, so i renamed every excel sheet manually



All of the new excel sheets with renamed column headers were then saved and then renamed themselves into a new excel file.

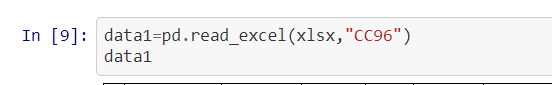


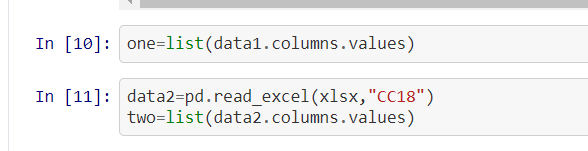
Cleaning\_up\_datasets\_second\_try.ipynb:

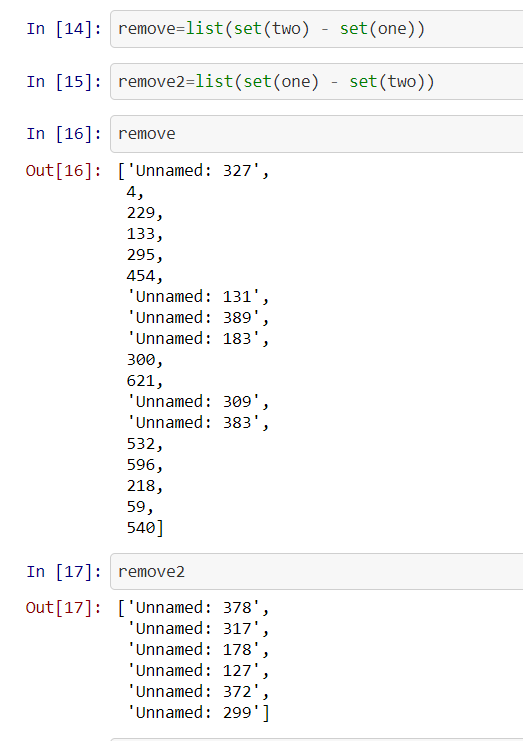
Using the new excel file we named “output”, noticed there were differences in column names from 1996-2011 and 2012-2018



Used the 1996 (CC96) and 2018 (CC18) files, made a list of column names, and then compared both lists to each other using python to note if one had column names that the other did not.

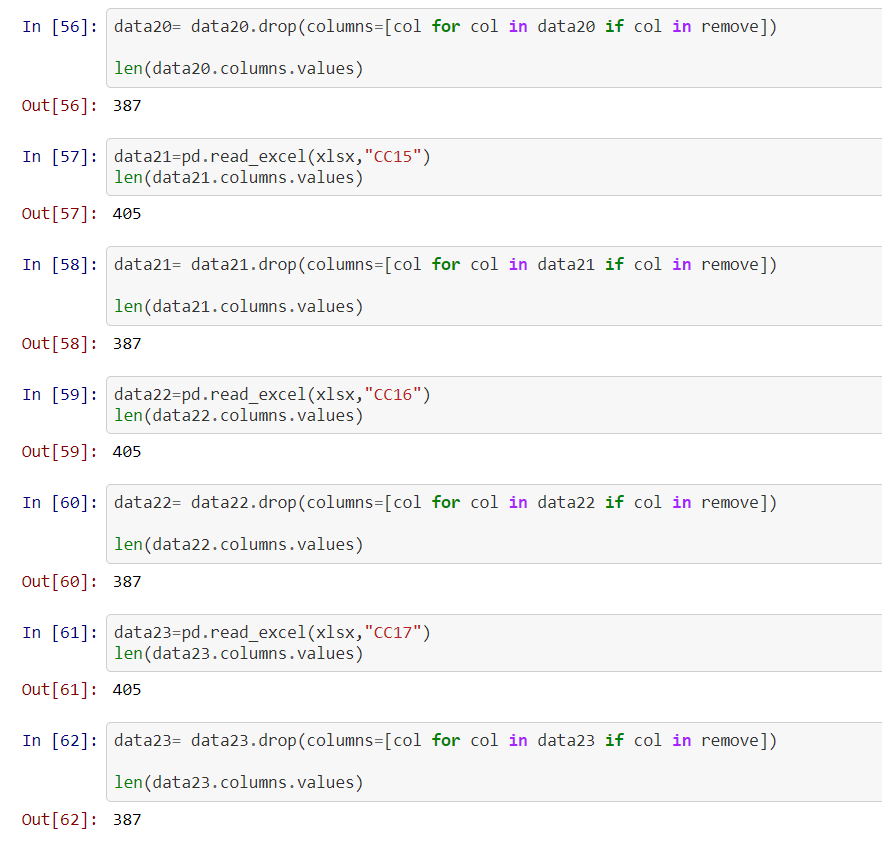






Data from 1996-2011 (CC96 to CC11) would need to have column names listed in “remove2” to be removed, while 2012-2018(CC12-CC18) would need names from “remove” removed.

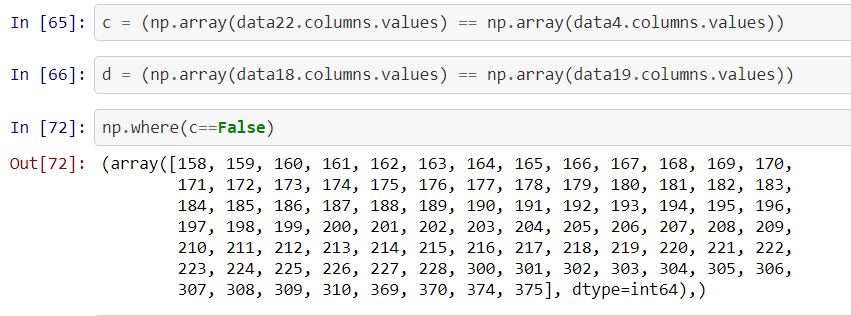




Ultimately, each excel sheet would have the same number of columns (387 total).

Still, noticed that there were issues with the datasets and that some of the data was not organized exactly the same even with the same number of columns.

Another issue with column names between the 1996-2011 and 2012-2018 data was that everything was not ordered the same, and so the list of column names from the different years, when compared, would not be exactly the same.



Manually noted the issues with column names and then changed them in each excel sheet, or simply removed them for simplicity.

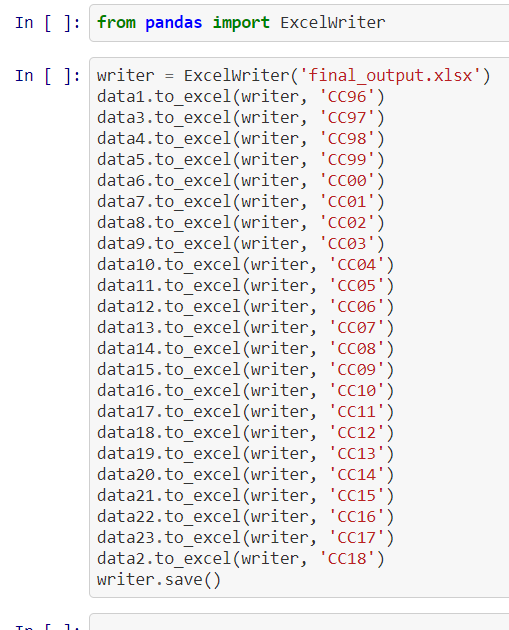
The phytoplankton were organized by group, and some phytoplankton species were placed in different groups over the years.

One example is Gonyaulax areolata, which was labeled differently over the years

(1996-2011 Thecate Dinoflaggelate

2012\_2018 Pennate Diatom)

With excel, I sorted within each group so that they were all organized alphabetically



Finalized and saved the final excel workbook that I completed.