10/20/2020 K-means clustering

K Means clustering was implemented on 2 datasets: dat\_means\_stock\_wide.csv and dat\_weights.csv. Each Google Colab notebook uses K-means for each dataset (with dat\_means\_stock\_gear still in progress). Afterwards, the clusters were visualized in the dimensionality reduction Colab notebook using PCA. After implementing K-means clustering on each dataset, a new dataset was saved with a new column for the clusters that each data point is assigned to (i.e., if there are 4 clusters, each data point would be assigned to a value 0 to 3 representing each clusters center). These new datasets are kmeans\_dat\_stock\_wide.csv and kmeans\_dat\_weights.csv

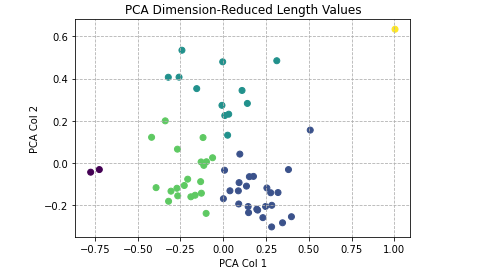
Results: K-means proved to work best with 4 clusters for the first dataset (dat\_means\_stock\_wide.csv) and 5 clusters for the second dataset (dat\_weights.csv). For the dataset **dat\_means\_stock.csv,** K-means was run with 3 clusters. The Silhouette score technique was used, and the highest score signified the best number of clusters (graphs are shown in the Colab notebooks).

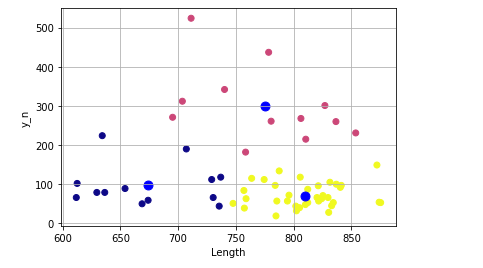
-The locations of the cluster centroids were also determined, with 4 centroids for **dat\_means\_stock\_wide.csv** data, 5 centroids for the **dat\_weights.csv dat**a, and 3 centroids for the **dat\_means\_stock.csv** data.

-PCA was applied to the final data, and the clusters were visualized in 2 dimensions with their distinct colors signifying their clusters.

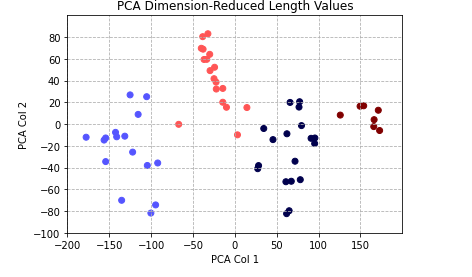
Overall, there didn’t seem to be any clear patterns in the data when it was visualized. The clusters did not seem very apparent without the colors showing each cluster. However, 4 clusters seemed to almost suit the first two datasets: dat\_means\_stock\_wide.csv and dat\_weights.csv (one data point, potentially an outlier, stood as its own cluster for the dat\_weights.csv data).

Cluster Images: (The colors indicate each separate cluster)

dat\_means\_stock.csv: blue cluster centers- 3 clusters dat\_weights.csv: (plotted using PCA)-5 clusters (yellow cluster has 1 point) 



dat\_means\_stock\_wide.csv: (plotted using PCA)



Jacks comments:

Thanks for your work on this, it all looks great! sorry we couldn’t meet sooner!

It looks like you have done a really good job working with the data frames, implementing the K-means algorithm and calculating the model fitting diagnostics. I think the results from the dat\_means\_stock\_wide data set are going to be particularly useful ,and so I think it is worth focusing some additional attention there.

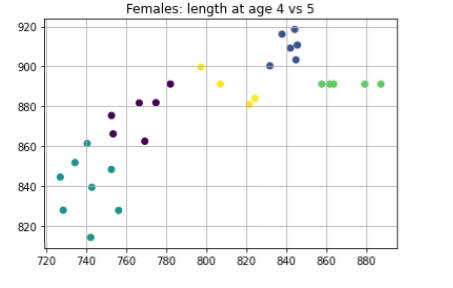
There are a few things that can be done moving forward that could be very useful or the project. The first would be to develop a method to help visualize the relationship between the clusters and biological factors such as run type, sex, and release type. This would help me understand the factors that might be causing the differences in length at age. We might also want to experiment a bit with the method used to impute the missing values. For example I expect that stocks of the same run type are more likely to be similar to each other than stocks of different run types, with this in mind we may want to impute missing sizes with the average for a run type rather than the mean for all stocks. Another thing that could help would be to group only by run type, release type, and rmis\_region, the different sexes of fish have different patterns of maturation so populations from one region are more likely to have five year old female fish than five year old males, the converse is true for three year old fish. This could help us get a better picture of the patterns of size at age caused by the different life history types and regions of origin.

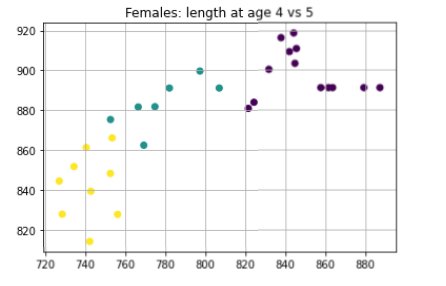
Once these two pieces are done I think we will have a pretty clear understanding of the time averaged patterns of size at age in the data set. And it would be great to move on towards trying to understand how these patterns are changing over time and if there are groups of stock with different temporal trends. Clustering based on trends over time will be a bit more difficult to analyze but I think there are several ways we could approach the problem. As a first pass I think we should focus just on the changes in size at age for four year old fish, because we have the most observations for this age class. We can then break period of time for which we have observations up into even sized bins (e.g. 1980 - 1989, 1990 - 1999, 2000-2010) and find the average length of four year old fish for each stock during those time blocks, we can then run k\_means using the averages for each time period as the variables of interest. It would be great to run this analysis for several different sets of time intervals and create some plots/tables that show the correspondence between the patterns we find and the biological variables like run type.

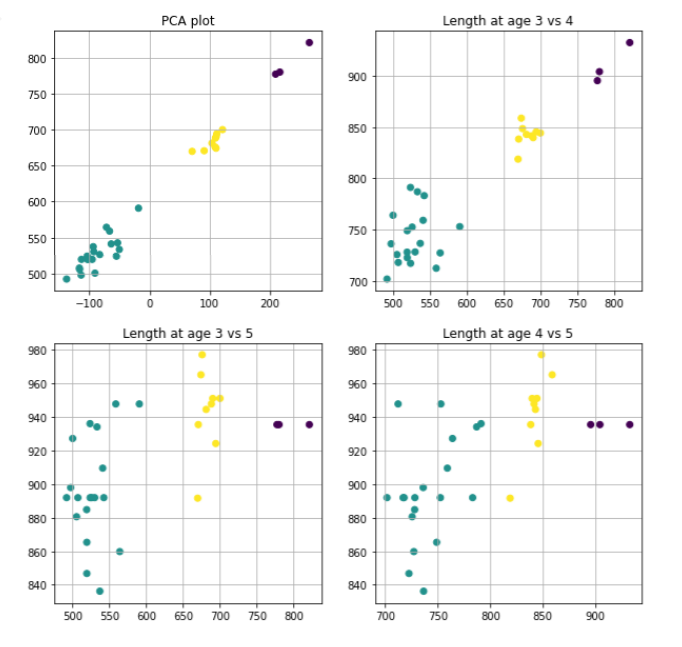
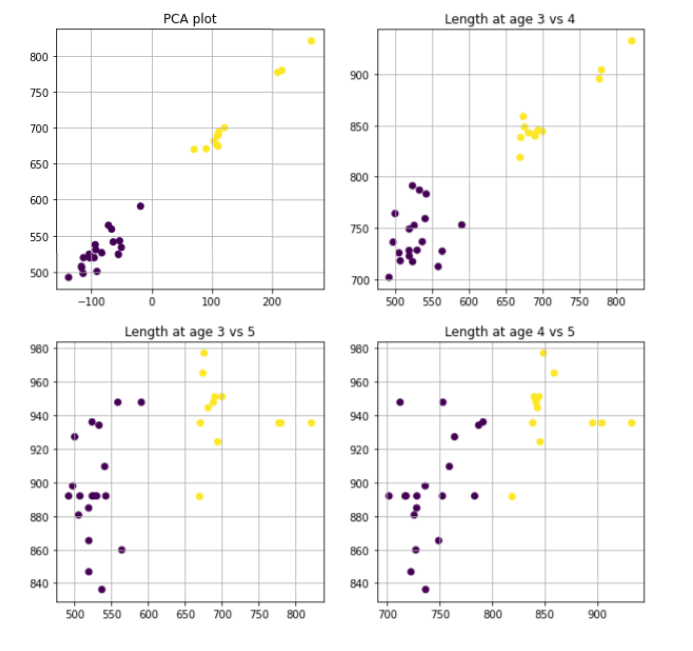
This approach is somewhat ad-hoc and it will be useful to develop a more formal method that handles missing value more effectively. To deal with some of these issues we can use likelihood based methods. For example we can fit a model that estimates K trends that each of the stocks follow over time, and then cluster the stocks based off of these trends using the expectation maximization algorithm. I will work on writing you a description of the technique and sending some readings so you can get up to speed on this if there is down time between when you finish the K\_means analysis and when we get to meet to discuss.

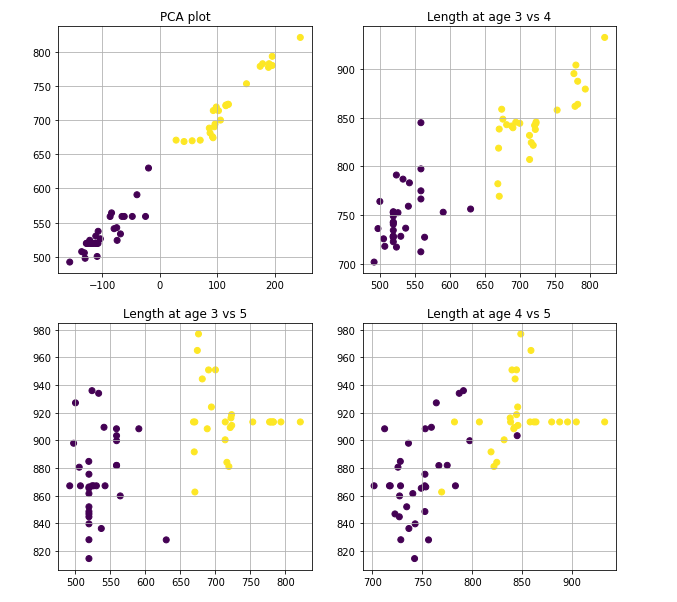
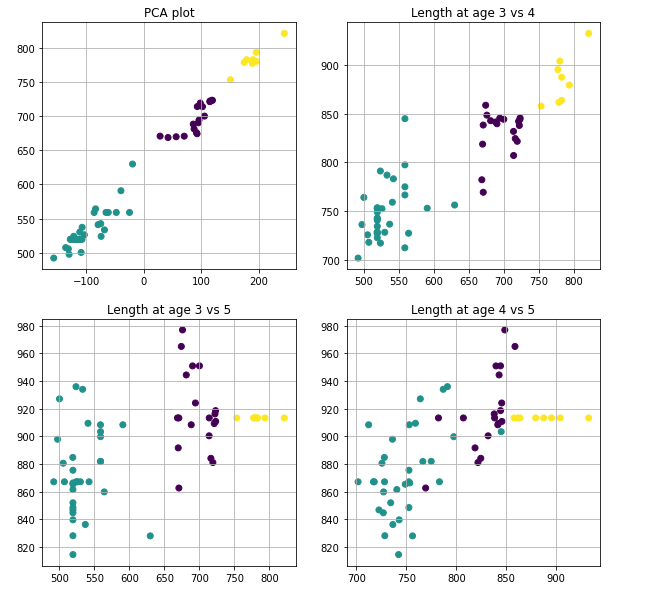
11/10/2020:

* K-means was implemented for 3 separate cases: females, males, and both. Each of these datasets were grouped based on **release\_type, run, and release\_location\_rmis\_basin** (as well as sex for the data including both females and males). These groups were averaged over the brood\_year and were placed back together into a single dataframe.
* The null values were imputed based on run type. Each set of data was grouped based on run, and the missing values in each group were replaced by the means for each group.
* K-means clustering was implemented on the imputed datasets using different numbers of clusters.

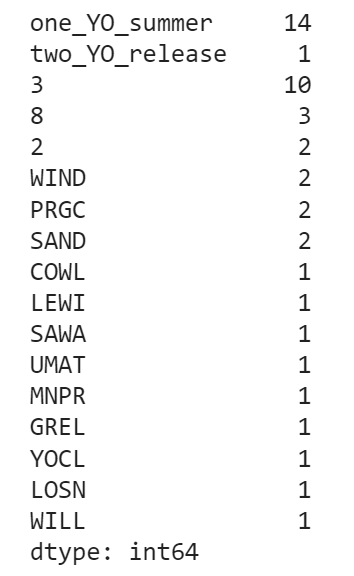
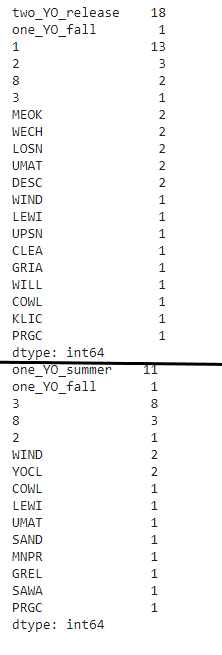
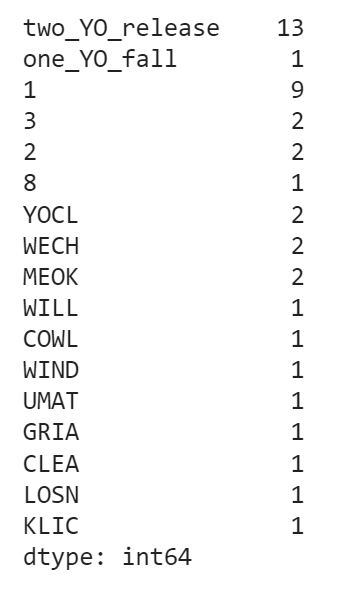
**Females: length at age 3 was disregarded due to remaining null values even after imputing**



**Males: PCA used**

**Both Males and Females:** 

**Categorical Data and Frequencies in each cluster:**

Females: Males: cluster 1 frequencies on top, cluster 2 frequencies on bottom

(e.g., 14 one\_YO\_summer release types in cluster 1)

Both: female and male categorical frequencies for clusters 1 and 2