Marine Animal Distribution in the California Current Ecosystem
Milestone 2

Anna Harvey
Spring 2021
https://github.com/anhar421/Portfolio

Which Domain?
There have not been any domain-specific surprises from the data so far. The team doing these yearly studies work out of Seattle in association with the University of Washington where I did my undergraduate degree in Aquatic and Fisheries Sciences. Often within that department, student have the opportunity to assist professors with research or observe or participate in other local fisheries sciences work. With that in mind, many of the concepts presented in this study are familiar to me. The interesting focus of the study is on using birds as indicator species for the overall ecosystem health. Although that is not the focus of my project, it is an interesting aspect to consider. I did not spend much time studying marine birds during my undergraduate program.

Which Data?
I spent this week primarily on EDA. I cleaned up my dataset and did a few rough visualizations to understand the data better. I figured out that the stations that they are tracking species near arrange themselves along the coast between Washington and Northern California in a fairly even distribution. I found out that there are 108 unique animal/object codes in the data that have been tracked. About half of the codes match bird codes from the American Ornithological Society codes (https://www.birdpop.org/pages/birdSpeciesCodes.php). The other half I will need more information in order to decode. I have contacted the data manager for the project in hopes of getting a codebook for the species codes as well as the station codes.

Research Questions? Benefits? Why analyze these data?
Since the stations are somewhat evenly distributed along the coast and fairly close together, I do not think I will be able to visually represent the species on a map in a way that is as useful as I thought it would be. I think there will still be some benefits to mapping out specific species over time or comparing a handful of species instead of all 108 groups represented in the data. A time series visualization of species at the stations may also be helpful. The dataset had no NaNs, so that will be helpful for predictive analytics. If I cannot get information on the species codes, I will choose to focus just on the identified bird species.

What Method?
I need to fine-tune my visualization strategy. Using a map would be both interesting and helpful considering that the focus of the study is to see if bird species distribution relates to salmonid distribution. Since there are 108 groups being tracked at each station, trying to visualize all of them at once is too crowded, even though not every group shows up at every station. I also need to figure out which predictive model to use. Since the goal is to try and predict if a species will show up at a specific station, I may be able to use clustering, time series, or forecast models.

Potential Issues?

Determining which predictive model to attempt will be my biggest hurdle along with actually executing it. I am hoping to get some feedback from my peers on which model would work the best. The other big issue is not having the codes yet for the species and stations. The station codes are not critical, as I can figure out which is which manually. Some of the unknown species codes are intuitive in nature (ie SHRK probably means “shark”). Since the goal of the project is to predict the locations of these groups, I could continue with them without having the codes and just keep the species codes as the name of the group. However, a more productive solution to that and the crowding of my visuals would be to just focus on a small handful of the bird species instead of all 108 groups.