Capstone 1: Data Wrangling Update and Data Story

Burns, Echelle

Dec 2019

Data wrangling efforts were previously described in the Data Wrangling section. However, due to the high abundance of zeros in the dataset, I performed a jackknifing technique to produce a dataset with a relatively equal number of data points between times when sharks were present and when sharks were absent. Zone IDs in which sharks were detected each year were gathered and a list of nearby zones (± 10 IDs away) was generated. A random sample of data with no sharks detected (non-shark dataset) was then taken. The random sample only collected data from zones that were within the zone list that was generated. The size of random samples varied across years, but were equal to the number of rows in the dataset in which sharks were present. Sometimes, the non-shark dataset with zones within the zone list was not large enough to perform the random sampling. In this case, all data within the non-shark dataset with relevant zones were gathered, and the remaining rows of data were randomly grabbed from the non-shark dataset with non-relevant zones. Making sure the data were gathered across similar zones was important to reduce biases that may arise from comparing shark data that primarily occurs in nearshore waters with non-shark data that may extend several hundred miles offshore.

In addition, the dataset was modified to answer a different question than the original dataset was designed to answer. Originally, the master dataset included data on individual sharks, such as sex, tagging location, and tagging cohort. However, for the purpose of this project, I will begin by asking whether environmental conditions influences the number of sharks that are detected in a particular grid cell. Unfortunately, this reduces my number of categorical data values to one, because most categorical data fields included individual shark data, such as those listed above.

Once jackknifed, all remaining data were visually assessed for outliers and descriptive statistics were conducted. For continuous and numeric datasets, histograms were plotted and the mean and all four quartiles were calculated. For categorical data (lunar phase), a barplot was created and the counts in each category were calculated. For data that were associated with time (number of sharks detected, receiver density, Sea Surface Temperature, Sea Surface Salinity, Chlorophyll-A), time series plots were also created in order to see annual and seasonal trends. In general, it appears that the number of sharks that are detected may rely quite heavily on the number of receivers that are in the water listening for tagged animals, as it is impossible to detect an individual when a receiver is not present. Of the environmental parameters, it appears that Sea Surface Temperature may also be driving shark presence, as has been indicated by previous hypotheses in elasmobranch research. There were outliers present in Chlorophyll-A (where most values were <1 but some values exceeded 50) and in the lunar phase (where some values were represented less frequently due to the inherent characteristics of the lunar cycle). All other values appeared to maintain a regular pattern across years and months with some variation.

Specific questions asked by SpringBoard in Section 7.2 can be found using 2017 data prior to jackkniffing (Data_Story Jupyter Notebook), whereas jackknifing techniques and

subsequent data visualization on the new dataset can be found in the Jackknifing Jupyter Notebook and the Distributions Jupyter Notebook, respectively.