Capstone 1: Data Wrangling
Burns, Echelle
Nov 2019

A receiver deployment log was provided by the CSULB Shark Lab. This file included information on when and where acoustic receivers were in the water, with parameters such as Receiver SN (serial number), Latitude, Longitude, DateTime In (local time), DateTime Out (local time), and Station Name. Receivers that were lost or still deployed in the water (receivers with NAs in either time column) were removed from the dataset and the datetimes were converted to UTC. The geospatial range from this dataset was calculated and used to gather environmental datasets from NOAA CoastWatch.

Acoustic receiver detection data were also provided by the CSULB Shark Lab and included fields such as Receiver ID, Transmitter Code, DateTime (UTC), Latitude, and Longitude. These data were pre-processed by the Lab to include only individuals that were detected by an acoustic receiver two or more times per day. This dataset was cleaned to only include data during times that receivers were known to be in the water, in order to reduce the probability of false detection, when a receiver suggests an animal was present when it really was not. In addition, only sharks that were on the CSULB Shark Lab's tag deployment datasheet were included in the final dataset. The tag deployment datasheet was merged with detection data to provide information on the tagging cohort (year) of the individual, as well as its weight (kg) at tagging, total length (cm) at tagging, sex, and tagging location (city).

Daily environmental datasets for sea surface temperature, chlorophyll-A and sea surface salinity, and overall seafloor depth gradient data were gathered from NOAA CoastWatch's ERDDAP and were downloaded to a local server. Each of these environmental datasets have been hypothesized to influence the behavior of other marine species. Due to the large spatial extent of these datasets, data were saved in one-year increments. Moon phases were also calculated using the pylunar package.

A shapefile of grid cells (0.01 x 0.01 decimal degrees) was used to standardize and merge the environmental datasets with the shark detection data. The grid ID number over which individual points were located was saved to each dataset. This allowed for the comparison of data across grid cells, as opposed to across individual latitude/longitude combinations. Several environmental datasets did not extend far enough into shallow waters to overlap with grid cells that contained receivers or shark detections. Therefore, missing data were filled using values from the nearest grid cell for that parameter during that day. In order to accomplish such filling, the data had to be ordered by date, latitude, and longitude. However, each individual dataset had its own unique set of latitude and longitude values. Therefore, the mean latitude and longitude for each zone were calculated as the average latitude and longitude across all geospatially-referenced datasets. Additionally, in order to calculate moon phase, latitude and longitude coordinates had to first be converted from decimal degrees to tuples of (degrees, minutes, seconds) and datetime had to first be converted to tuples of (year, month, day, hour, minute, second). The appropriate conversions were conducted before pylunar was called.

The final dataset included the fields: Zone (grid cell), Date, Transmitter ID, weight, length, sex, tagging location, tagging cohort, receiver density (calculated for each day in each

grid cell as the number of receivers deployed per cell), latitude, longitude, seafloor depth gradient, sea surface temperature, sea surface salinity, chlorophyll-A, and moon phase. NAs will remain in the dataset for Transmitter ID, weight, length, sex, tagging location, tagging cohort to indicate a date and zone in which no sharks were present.

The Jupyter Notebooks used for this project were run in the following order: Preliminary_Data_Wrangling, Data_Wrangling_Shapefiles, Building_Environmental_Indices, Merging_Environmental_Data, Merging_To_Master_Datasheets, and Adding_Moon_Phases. All notebooks can be found in my GitHub Repository under the Data_Wrangling folder.