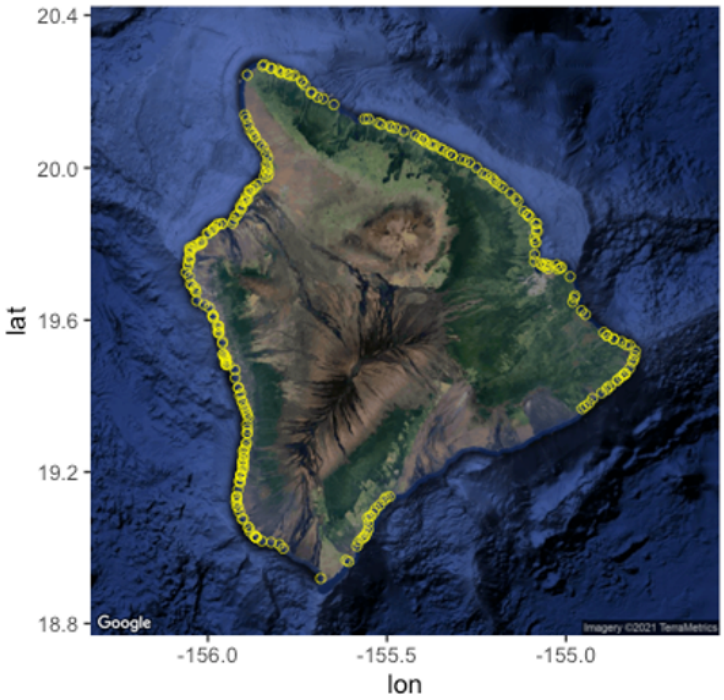
Homework 3 – Some Programming Practice



The goal of this assignment is to have you practice some basic R programming / data wrangling. The attached file, "CRCP_Reef_Fish_Surveys_Hawaii_expanded.csv", includes survey data of reef fish on Hawai'i island collected by NOAA (yellow circles in the map above = survey location, see here). I have slightly modified the data (added zeros for when a species it not observed at a location, removed some columns we will not be using).

(1) Let's look at some basic summary statistics. In total there are 209 fish species encountered and counted in these surveys. Identify the top twelve species in terms of these three statistics: (1) mean abundance, (2) maximum abundance, (3) standard deviation of abundance. In other words, which species are most abundant on average, which reach the highest abundances, and which have the most variable abundances?

To make these calculations, you'll want to focus on the 'count' column (number of fish observed at a location), as well as the 'commonname' or 'taxonname' columns. The latter two are nearly identical, you can choose which you prefer to use. To do these calculations in base R, tapply() or aggregate() will be useful, or to use the 'tidyverse' you can look into the summarize() function in the package dplyr.

(2) Now let's visualize how the abundances of the most common species vary with depth, to get a sense for whether species have different depth niches. Using the top twelve species based on mean abundance, plot a scatterplot of count vs. depth (the column is named 'depth'). Add a smoother to help visualize mean count vs. depth, put all twelve

species on a single plot (with twelve panels), and make sure each panel is titled using the species name. Furthermore, do not arrange the panels in alphabetical order — instead, arrange them in order of mean abundance, so that the most abundant species is in the top left panel, and remaining panels are in order of mean abundance. The patterns in abundance will be easier to see if you transform the counts or the y-axis, for example with a square root, because the counts are very skewed.

I would like you to make this twelve-paneled figure in two ways. First, write a 'for' loop in which you iterate your plotting code for the twelve species. The function scatter.smooth() is a simple one for making a scatterplot with a smoother. Second, use ggplot() to do the same thing. The functions geom_point(), geom_smooth(), and facet_wrap() will be helpful.

The ggplot way of doing this may be slightly easier (once you've figured out the ggplot syntax), but I would like you to do it both ways so that I know that you know how to write a loop, which is a very general and important programming technique.

What are your conclusions from this visual inspection of the data? Make sure that the patterns in the plots are visible. If they are not, you may need to adjust the figure dimensions in Markdown.

(3) Finally, let's focus on just the top 5 most abundant species. Make a new plot that shows abundance vs. depth for the top 5 species, including smoothers, but this time put all of the species on the same scatterplot and distinguish them with different colors. This time you don't need to use two different approaches to make the plot — one approach will suffice. What is your interpretation of this plot?

Fit a linear model that tests whether these 5 species have different depth niches. Based on residual diagnostic plots of this model, and a plot of the fitted effects, do you think a linear model is a good approach for testing this question? If not, why not?