Paige Duffin Data Analysis Project

Working title- will give proper title later

Paige Duffin

2019-10-11

# Organizing this into tabs for the time being

## Part 1: Introduction

Summary/Abstract *Write a summary of your project.*

### General Background Information

Outbreaks of infectious disease are an important, yet poorly understood, driving force in population biology and, while prevalent in both terrestrial and marine systems, the manner by which outbreaks influence biotic relationships, age demographics, community structure and function, and trophic interactions in the aquatic realm consistently lags behind that of terrestrial disease ecology. The growing body of evidence in this field, however, supports the observation that marine disease epidemics are increasing in frequency and severity (Orth et al., 2006; Waycott et al., 2009). Elucidating the relationship between aquatic species and their pathogenic diseases is pertinent in the field of marine ecology because infection outbreaks have the potential to drastically alter ecosystem functionality (Blakesley et al., 2002; Burge et al., 2013). Many of the organisms which have faced chronic and/or severe outbreaks of disease (such as sea urchins, scleractinian corals and seagrasses) are also considered keystone species and/or ecosystem engineers, meaning that they contribute highly valuable services or functions to their surrounding community and ecosystem. Thus, it comes as no surprise that disease-driven mass mortalities of these species often generate waves of ecological permutation, ranging from temporary local disruptions to permanent phase shifts (Burge et al., 2013), such as the transformation from a coral to macroalgal dominated Caribbean reef structure that accompanied the massive *Diadema antillarum* (black sea urchin) die-off of the early 1980s (Lessios et al., 1984; Lessios, 1988). This event, likely caused by a biological pathogen (Schultz et al., 2016), reduced Diadema populations to less than 1% of their original size (Lessios, 1988). Mass mortality events impacting critical foundation species, ecosystem engineers, or keystone species such as the black sea urchin have been coined ‘marine disease emergencies’ due to the detrimental cascade of events that often succeeds them (Miner et al., 2018).

Sea Star Wasting Disease (SSWD) is another epizootic crisis facing modern day coastal ecosystems. Outbreaks of asteroid wasting have been documented periodically since the late 1970s and SSWD describes a suite of symptoms observed across a broad range of sea star species, most of which play integral parts in shaping their community structure. Generally speaking, the wasting disease events which punctuated the past four decades were relatively brief, in localized areas, and largely failed to capture the attention of the scientific community. Beginning in summer 2013, however, mass mortalities of sea stars due to wasting disease have caused unprecedented damage, owing to the geographical and temporal extent of impact. This ongoing epizootic, which has killed millions of asteroids across over 20 taxa, is widely referred to as the largest disease event sweeping through a wildlife marine species in documented history (Hewson et al. 2014, Gudenkauf & Hewson, 2015, Eisenlord et al. 2016).

### Description of data and data source

*Describe what the data is, what it contains, where it is from, etc.*

#### The Source

The vast majority of disease documentation and surveys cover the west coast of the United States. While it is almost certainly true that SSWD is most prevalent in this region, there is a strong bias towards the discovery of diseased organisms owing to the emphasis of long-term survey networks in the area, such as the extensive [Multi-Agency Rocky Intertidal Network](https://marine.ucsc.edu/index.html) (MARINe for short). The vast majority of what we know about SSWD dynamics comes from studies along the Pacific Coast of North America using data produced through the combined effort of dedicated scientists, government agencies, and local citizens. The purpose of this investigation is to make use of this data monitoring system to investigate spatial and temoporal changes in sea star abundance along the west coast of the United States.

#### The Data

Below I explore the raw data I downloaded from a link I recieved after submitting a request for the data on the MARINe webpage.

library(readr)  
raw\_SS\_count <- read.csv("~/Documents/GitHub/PaigeDuffin-Project/data/raw\_data/seastarkat\_size\_count\_totals\_download.csv")  
head(raw\_SS\_count, 3)

## group\_code site\_code marine\_site\_name marine\_sort\_order latitude  
## 1 UCLA ALEG Alegria 6420 34.46714  
## 2 UCLA ALEG Alegria 6420 34.46714  
## 3 UCLA ALEG Alegria 6420 34.46714  
## longitude marine\_common\_season marine\_season\_code marine\_common\_year  
## 1 -120.2774 85 SP02 2002  
## 2 -120.2774 85 SP02 2002  
## 3 -120.2774 85 SP02 2002  
## season\_sequence season\_name target\_assemblage method\_code species\_code  
## 1 1 Spring sea\_star IP PISOCH  
## 2 1 Spring sea\_star IP PISOCH  
## 3 1 Spring sea\_star IP PISOCH  
## size\_sort\_order size\_bin total mpa\_designation mpa\_region georegion  
## 1 3 20 1 reference South Coast CA South  
## 2 4 30 3 reference South Coast CA South  
## 3 5 40 16 reference South Coast CA South  
## bioregion state\_province island last\_updated  
## 1 Government Point to Mexico California Mainland 2019-04-05 20:07:59  
## 2 Government Point to Mexico California Mainland 2019-04-05 20:07:59  
## 3 Government Point to Mexico California Mainland 2019-04-05 20:07:59

levels(raw\_SS\_count$georegion)

## [1] "AK" "CA Central"   
## [3] "CA Channel Islands South" "CA North"   
## [5] "CA North Central" "CA South"   
## [7] "OR" "WA Olympic Coast"   
## [9] "WA Salish Sea"

levels(raw\_SS\_count$species\_code)

## [1] "EVATRO" "KATTUN" "PISOCH"

Notably, it contains the following information (some parameters removed from this overview):

* groupcode: The unique code for each monitoring group
* marine\_site\_name: The name of the site where the survey was conducted.
* latitude and longitude: (self explanatory)
* marine\_season\_code: A four-character code to identify the Sampling season. The first two characters indicate the season and the last 2 characters indicate the year.
* method\_code: The method used for sampling the plot.
* size\_bin: The size of the species being counted, binned to the nearest 5 or 10 millimeter.
* target\_assemblage: The specific organism that this transect was created to monitor.
* total: Total number of individuals counted in a given size\_bin
* mpa\_designation: Describes whether the referenced site is located within a Marine Protected Area (MPA) or is a reference site. If this field is blank, the referenced site is not located within an MPA, or is not a reference site for an existing MPA site.
* georegion/ bioregion: geographic/biogeographic region in which site is located
* state\_province: The State or Province and Country where the referenced site is located.
* island: The name of the island where the referenced site is located. Sites not on islands are designated as mainland.

#### Additional Data Sources

I believe I have more data coming in to work with. I recieved an email today from a representative of the Multi-Agency Rocky Intertidal Network. This email said:

*“You should have been able to download our sea star size/count data from the data download page that you were directed to after filling out the data request form. We also have associated disease category data, but because we only sample 1x per year at most sites, this additional info will not give you an idea of disease emergence or progression within a site or region. You might also be interested in our observation data set, which would be better for these types of questions. We can have this data to you within about a week.”*

I’m also interested in environmental data such as sea surface temperature, dissolved oxygen, salinity, etc., as SSWD is thought to be exacerbated by abiotic stressors. However, I’m having a difficult time finding data to use that will pair reasonably well with the dates and locations of the raw\_SS\_count data I know I will be working with.

I’m very interested in this disease and environmental data, but don’t know what it will look like or if it will be feasible to use. Hopefully, I can incorporate this into my analysis, but for right now I am going to focus on outlining the questions I can answer with the data I have in hand.

### Questions/Hypotheses to be addressed

*State the research questions you plan to answer with this analysis*

**1. Are sea star populations changing over time?** I anticipate that there will be a decline in sea star populations over time, but that the change will not occur uniformly across geographical regions and species. Species that are reported to have been affected by SSWD will experience the biggest declines.

**2. Is species composition uniform across space?** I expect to be able to define rough species ranges.

**3. Is there evidence of sampling bias?** Is there a relationship between important survey parameters (mainly, abundance data) and sampling method (ex. group conducting the survey, plot sampling method, the target assemblage). At the very least, I expect there will be bias based on the targeted organism of the survey; if a given survey aims to characterize the presence of one species of sea stars, it is likely that they will seek out that species and record higher counts.

#### Questions I may be able to answer with disease and/or environmental data:

**1. Is there a relationship between sea star abundance over time and prevalence of SSWD?** There absolutely should be a relationship, based on the body of evidence presented on studies that have focused on individual species.

**2. Are there species of sea stars that are more affected by disease than others?** The literature implies that this may be the case, but there also may be a bias based on species that researchers are more interested in.

**3. Are there regions that are more affected by disease than others, regardless of species?** The most recent epidemic appears to be so widespread that I suspect there won’t be any significant trends here, but it’s certainly worth checking.

**4. Is there a relationship between sea star wasting disease and dissolved oxygen levels?** A recent, but leading hypothesis in the field of SSWD suggests that it may be triggered specifically by low oxygen conditions. I’m very interested in exploring relationships between SSWD and DO levels- this is actually the main curiosity that lead me to focus in on this data set.

**5. Is there a relationship between sea star wasting disease and [other water quality parameters I’m able to find data for]?**

## Part 2 Submission:

### Recreating a Visualization using R Code

I have spent a very long time trying to figure out how to get all of the different components of the .Rproj to “talk” to each other (ie, write script in “processingscript.R”, have it save the processed script to the ./data/processed\_data/ folder as an “RDS”) and I can’t figure it out. I looked at the example provided by Dr. Handel/his student in the second module, as well as some of the other students who have submitted already publically on the GitHub page.

In the interest of having something to turn in for this assignment on time, I am going to make this R Markdown file on my progress and turn it into a website, as this is something I do know how to do. I don’t want to complain, but I have to admit I am feeling very frusterated, overwhelmed, and defeated right now :(

Throughout this document, I’ve highlighted places where I’ve had an especially difficult time in font color=“red”>red font.

font color=“red”>Generally speaking, I do not know how to use the bibtex reference approach. I may be able to figure it out on my own, but truly am worried about how much of a time commitment this is all going to be. Are there any resources you can point me towards so I can learn how to use this system?

### Processing Script

This is what I would include in my processingscript.R file.

processing script

this script loads the raw data, processes and cleans it and saves it as Rds file in the processed\_data folder

load needed packages. make sure they are installed.

library(caret) # Classification and Regression Training

## Warning: package 'caret' was built under R version 3.5.2

## Warning: package 'ggplot2' was built under R version 3.5.2

library(corrplot)  
library(dplyr) # A Grammar of Data Manipulations

## Warning: package 'dplyr' was built under R version 3.5.2

library(forcats) # Tools for Working with Categorical Variables (Factors)

## Warning: package 'forcats' was built under R version 3.5.2

library(gdata) # Various R Programming Tools for Data Manipulation  
library(ggplot2) # Create Elegant Data Visualizations Using the Grammar of Graphics  
library(ggrepel) # Automatically Position Non-Overlapping Text Labels with ‘ggplot2’

## Warning: package 'ggrepel' was built under R version 3.5.2

library(ggridges) # Ridgeline Plots in ‘ggplot2’  
library(ggthemes) # Extra Themes, Scales, and Geoms for ‘ggplot2’

## Warning: package 'ggthemes' was built under R version 3.5.2

library(gmodels) # Various R Programming Tools for Model Fitting  
library(grid) # The Grid Graphics Package  
library(gridExtra)  
library(knitr) # A General-Purpose Package for Dynamic Report Generation in R

## Warning: package 'knitr' was built under R version 3.5.2

library(lmtest) # Testing Linear Regression Models

## Warning: package 'lmtest' was built under R version 3.5.2

library(plyr) # Tools for Splitting, Applying and Combining Data  
library(purrr) # Functional Programming Tools

## Warning: package 'purrr' was built under R version 3.5.2

library(readr) # Read Rectangular Text Data  
library(rpart)

## Warning: package 'rpart' was built under R version 3.5.2

library(rattle)  
library(scales) # Scale Functions for Visualization  
library(shiny) # Web Application Framework for R

## Warning: package 'shiny' was built under R version 3.5.2

library(skimr) # Compact and Flexible Summaries of Data

## Warning: package 'skimr' was built under R version 3.5.2

library(stringr) # Simple, Consistent Wrappers for Common String Operations  
library(tibble) # Simple Data Frames

## Warning: package 'tibble' was built under R version 3.5.2

library(tidyr) # Easily Tidy Data with ‘spread()’ and ‘gather()’ Functions

## Warning: package 'tidyr' was built under R version 3.5.2

library(tidyverse) # Install and Load ‘Tidyverse’ Packages  
library(usmap)

## Warning: package 'usmap' was built under R version 3.5.2

library(rworldmap)  
library(ggmap)

## Warning: package 'ggmap' was built under R version 3.5.2

library(maps)  
library(mapdata)  
library(cowplot)

## Warning: package 'cowplot' was built under R version 3.5.2

library("googleway")  
library("ggspatial")   
library("rnaturalearth")   
library("rnaturalearthdata")  
theme\_set(theme\_bw())  
library("sf")

## Warning: package 'sf' was built under R version 3.5.2

register\_google(key = "AIzaSyB3n0SkJTXxrbUkF2bkBJs7XdleT\_iJXsg", write = TRUE)  
library(readxl)

## Warning: package 'readxl' was built under R version 3.5.2

library(dplyr)  
library(readr)

load data. path is relative to project directory.

# raw\_SS\_count <- read.csv("./data/raw\_data/seastarkat\_size\_count\_totals\_download.csv") # already did that, above  
head(raw\_SS\_count, 3)

## group\_code site\_code marine\_site\_name marine\_sort\_order latitude  
## 1 UCLA ALEG Alegria 6420 34.46714  
## 2 UCLA ALEG Alegria 6420 34.46714  
## 3 UCLA ALEG Alegria 6420 34.46714  
## longitude marine\_common\_season marine\_season\_code marine\_common\_year  
## 1 -120.2774 85 SP02 2002  
## 2 -120.2774 85 SP02 2002  
## 3 -120.2774 85 SP02 2002  
## season\_sequence season\_name target\_assemblage method\_code species\_code  
## 1 1 Spring sea\_star IP PISOCH  
## 2 1 Spring sea\_star IP PISOCH  
## 3 1 Spring sea\_star IP PISOCH  
## size\_sort\_order size\_bin total mpa\_designation mpa\_region georegion  
## 1 3 20 1 reference South Coast CA South  
## 2 4 30 3 reference South Coast CA South  
## 3 5 40 16 reference South Coast CA South  
## bioregion state\_province island last\_updated  
## 1 Government Point to Mexico California Mainland 2019-04-05 20:07:59  
## 2 Government Point to Mexico California Mainland 2019-04-05 20:07:59  
## 3 Government Point to Mexico California Mainland 2019-04-05 20:07:59

levels(raw\_SS\_count$georegion)

## [1] "AK" "CA Central"   
## [3] "CA Channel Islands South" "CA North"   
## [5] "CA North Central" "CA South"   
## [7] "OR" "WA Olympic Coast"   
## [9] "WA Salish Sea"

levels(raw\_SS\_count$species\_code)

## [1] "EVATRO" "KATTUN" "PISOCH"

take a look at the data

dplyr::glimpse(raw\_SS\_count)

## Observations: 13,165  
## Variables: 24  
## $ group\_code <fct> UCLA, UCLA, UCLA, UCLA, UCLA, UCLA, UCLA, U…  
## $ site\_code <fct> ALEG, ALEG, ALEG, ALEG, ALEG, ALEG, ALEG, A…  
## $ marine\_site\_name <fct> Alegria, Alegria, Alegria, Alegria, Alegria…  
## $ marine\_sort\_order <int> 6420, 6420, 6420, 6420, 6420, 6420, 6420, 6…  
## $ latitude <dbl> 34.46714, 34.46714, 34.46714, 34.46714, 34.…  
## $ longitude <dbl> -120.2774, -120.2774, -120.2774, -120.2774,…  
## $ marine\_common\_season <int> 85, 85, 85, 85, 85, 85, 85, 85, 85, 85, 85,…  
## $ marine\_season\_code <fct> SP02, SP02, SP02, SP02, SP02, SP02, SP02, S…  
## $ marine\_common\_year <int> 2002, 2002, 2002, 2002, 2002, 2002, 2002, 2…  
## $ season\_sequence <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…  
## $ season\_name <fct> Spring, Spring, Spring, Spring, Spring, Spr…  
## $ target\_assemblage <fct> sea\_star, sea\_star, sea\_star, sea\_star, sea…  
## $ method\_code <fct> IP, IP, IP, IP, IP, IP, IP, IP, IP, IP, IP,…  
## $ species\_code <fct> PISOCH, PISOCH, PISOCH, PISOCH, PISOCH, PIS…  
## $ size\_sort\_order <int> 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 15, 16…  
## $ size\_bin <int> 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 1…  
## $ total <int> 1, 3, 16, 25, 27, 22, 11, 2, 2, 1, 1, 3, 6,…  
## $ mpa\_designation <fct> reference, reference, reference, reference,…  
## $ mpa\_region <fct> South Coast, South Coast, South Coast, Sout…  
## $ georegion <fct> CA South, CA South, CA South, CA South, CA …  
## $ bioregion <fct> Government Point to Mexico, Government Poin…  
## $ state\_province <fct> California, California, California, Califor…  
## $ island <fct> Mainland, Mainland, Mainland, Mainland, Mai…  
## $ last\_updated <fct> 2019-04-05 20:07:59, 2019-04-05 20:07:59, 2…

head(raw\_SS\_count, 3)

## group\_code site\_code marine\_site\_name marine\_sort\_order latitude  
## 1 UCLA ALEG Alegria 6420 34.46714  
## 2 UCLA ALEG Alegria 6420 34.46714  
## 3 UCLA ALEG Alegria 6420 34.46714  
## longitude marine\_common\_season marine\_season\_code marine\_common\_year  
## 1 -120.2774 85 SP02 2002  
## 2 -120.2774 85 SP02 2002  
## 3 -120.2774 85 SP02 2002  
## season\_sequence season\_name target\_assemblage method\_code species\_code  
## 1 1 Spring sea\_star IP PISOCH  
## 2 1 Spring sea\_star IP PISOCH  
## 3 1 Spring sea\_star IP PISOCH  
## size\_sort\_order size\_bin total mpa\_designation mpa\_region georegion  
## 1 3 20 1 reference South Coast CA South  
## 2 4 30 3 reference South Coast CA South  
## 3 5 40 16 reference South Coast CA South  
## bioregion state\_province island last\_updated  
## 1 Government Point to Mexico California Mainland 2019-04-05 20:07:59  
## 2 Government Point to Mexico California Mainland 2019-04-05 20:07:59  
## 3 Government Point to Mexico California Mainland 2019-04-05 20:07:59

levels(raw\_SS\_count$georegion)

## [1] "AK" "CA Central"   
## [3] "CA Channel Islands South" "CA North"   
## [5] "CA North Central" "CA South"   
## [7] "OR" "WA Olympic Coast"   
## [9] "WA Salish Sea"

levels(raw\_SS\_count$species\_code)

## [1] "EVATRO" "KATTUN" "PISOCH"

change many of the variables from factors to characters, at least for now

SS\_count <- raw\_SS\_count  
SS\_count$group\_code <- as.character(SS\_count$group\_code)  
SS\_count$site\_code <- as.character(SS\_count$site\_code)  
SS\_count$marine\_site\_name <- as.character(SS\_count$marine\_site\_name)  
SS\_count$marine\_season\_code <- as.character(SS\_count$marine\_season\_code)  
SS\_count$season\_name <- as.character(SS\_count$season\_name)  
SS\_count$target\_assemblage <- as.character(SS\_count$target\_assemblage)  
SS\_count$method\_code <- as.character(SS\_count$method\_code)  
SS\_count$species\_code <- as.character(SS\_count$species\_code)  
SS\_count$mpa\_designation <- as.character(SS\_count$mpa\_designation)  
SS\_count$mpa\_region <- as.character(SS\_count$mpa\_region)  
SS\_count$georegion <- as.character(SS\_count$georegion)  
SS\_count$bioregion <- as.character(SS\_count$bioregion)  
SS\_count$state\_province <- as.character(SS\_count$state\_province)  
SS\_count$island <- as.character(SS\_count$island)  
SS\_count$last\_updated <- as.character(SS\_count$last\_updated)  
  
glimpse(SS\_count)

## Observations: 13,165  
## Variables: 24  
## $ group\_code <chr> "UCLA", "UCLA", "UCLA", "UCLA", "UCLA", "UC…  
## $ site\_code <chr> "ALEG", "ALEG", "ALEG", "ALEG", "ALEG", "AL…  
## $ marine\_site\_name <chr> "Alegria", "Alegria", "Alegria", "Alegria",…  
## $ marine\_sort\_order <int> 6420, 6420, 6420, 6420, 6420, 6420, 6420, 6…  
## $ latitude <dbl> 34.46714, 34.46714, 34.46714, 34.46714, 34.…  
## $ longitude <dbl> -120.2774, -120.2774, -120.2774, -120.2774,…  
## $ marine\_common\_season <int> 85, 85, 85, 85, 85, 85, 85, 85, 85, 85, 85,…  
## $ marine\_season\_code <chr> "SP02", "SP02", "SP02", "SP02", "SP02", "SP…  
## $ marine\_common\_year <int> 2002, 2002, 2002, 2002, 2002, 2002, 2002, 2…  
## $ season\_sequence <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…  
## $ season\_name <chr> "Spring", "Spring", "Spring", "Spring", "Sp…  
## $ target\_assemblage <chr> "sea\_star", "sea\_star", "sea\_star", "sea\_st…  
## $ method\_code <chr> "IP", "IP", "IP", "IP", "IP", "IP", "IP", "…  
## $ species\_code <chr> "PISOCH", "PISOCH", "PISOCH", "PISOCH", "PI…  
## $ size\_sort\_order <int> 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 15, 16…  
## $ size\_bin <int> 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 1…  
## $ total <int> 1, 3, 16, 25, 27, 22, 11, 2, 2, 1, 1, 3, 6,…  
## $ mpa\_designation <chr> "reference", "reference", "reference", "ref…  
## $ mpa\_region <chr> "South Coast", "South Coast", "South Coast"…  
## $ georegion <chr> "CA South", "CA South", "CA South", "CA Sou…  
## $ bioregion <chr> "Government Point to Mexico", "Government P…  
## $ state\_province <chr> "California", "California", "California", "…  
## $ island <chr> "Mainland", "Mainland", "Mainland", "Mainla…  
## $ last\_updated <chr> "2019-04-05 20:07:59", "2019-04-05 20:07:59…

skim(SS\_count)

## Skim summary statistics  
## n obs: 13165   
## n variables: 24   
##   
## ── Variable type:character ─────────────────────────────────────────────────────────  
## variable missing complete n min max empty n\_unique  
## bioregion 0 13165 13165 13 33 0 6  
## georegion 0 13165 13165 2 24 0 9  
## group\_code 0 13165 13165 2 6 0 9  
## island 0 13165 13165 3 9 0 7  
## last\_updated 0 13165 13165 19 19 0 1  
## marine\_season\_code 0 13165 13165 4 4 0 57  
## marine\_site\_name 0 13165 13165 5 42 0 77  
## method\_code 0 13165 13165 2 4 0 4  
## mpa\_designation 0 13165 13165 2 9 0 5  
## mpa\_region 0 13165 13165 2 19 0 6  
## season\_name 0 13165 13165 4 6 0 4  
## site\_code 0 13165 13165 3 4 0 77  
## species\_code 0 13165 13165 6 6 0 3  
## state\_province 0 13165 13165 6 10 0 4  
## target\_assemblage 0 13165 13165 8 8 0 1  
##   
## ── Variable type:integer ───────────────────────────────────────────────────────────  
## variable missing complete n mean sd p0 p25  
## marine\_common\_season 0 13165 13165 116.82 19.23 78 101  
## marine\_common\_year 0 13165 13165 2009.7 4.81 2000 2006  
## marine\_sort\_order 0 13165 13165 5566.24 1125.15 1720 5020  
## season\_sequence 0 13165 13165 2.03 0.81 1 1  
## size\_bin 0 13165 13165 82.44 42.2 5 50  
## size\_sort\_order 0 13165 13165 9.57 4.15 1 6  
## total 0 13165 13165 9.92 16.43 1 2  
## p50 p75 p100 hist  
## 118 131 152 ▅▅▆▇▇▇▇▅  
## 2010 2013 2018 ▃▅▅▇▇▆▆▆  
## 6090 6370 7650 ▁▁▁▂▃▃▇▁  
## 2 3 4 ▇▁▇▁▁▇▁▁  
## 80 110 320 ▅▇▇▃▁▁▁▁  
## 10 12 33 ▅▇▇▃▁▁▁▁  
## 4 11 360 ▇▁▁▁▁▁▁▁  
##   
## ── Variable type:numeric ───────────────────────────────────────────────────────────  
## variable missing complete n mean sd p0 p25 p50  
## latitude 0 13165 13165 38.67 5.2 32.87 34.73 36.62  
## longitude 0 13165 13165 -122.25 2.86 -135.38 -123.98 -121.94  
## p75 p100 hist  
## 41.65 57.05 ▇▆▃▂▁▁▁▁  
## -120.62 -117.25 ▁▁▁▁▅▇▆▃

save as RDS file

saveRDS(SS\_count, file = “./data/processed\_data/”)

font color=“red”>**I don’t understand this step. After much trial and error, it continues to not work for me. Can you provide additional instruction on how to do this step? I keep recieving the error message: Error in gzfile(file, mode) : cannot open the connection. In addition: Warning message: In gzfile(file, mode) : cannot open compressed file ‘./data/processed\_data/’, probable reason ‘Is a directory’, but when I write “./data/processed\_data/processed\_SS\_count”, it gives me a different error**

### Exploring Study Sites

load necessary packages

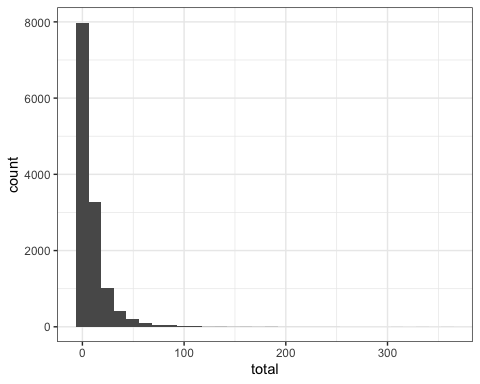
library(caret) # Classification and Regression Training  
library(corrplot)  
library(dplyr) # A Grammar of Data Manipulations  
library(forcats) # Tools for Working with Categorical Variables (Factors)  
library(gdata) # Various R Programming Tools for Data Manipulation  
library(ggplot2) # Create Elegant Data Visualizations Using the Grammar of Graphics  
library(ggrepel) # Automatically Position Non-Overlapping Text Labels with ‘ggplot2’  
library(ggridges) # Ridgeline Plots in ‘ggplot2’  
library(ggthemes) # Extra Themes, Scales, and Geoms for ‘ggplot2’  
library(gmodels) # Various R Programming Tools for Model Fitting  
library(grid) # The Grid Graphics Package  
library(gridExtra)  
library(knitr) # A General-Purpose Package for Dynamic Report Generation in R  
library(lmtest) # Testing Linear Regression Models  
library(plyr) # Tools for Splitting, Applying and Combining Data  
library(purrr) # Functional Programming Tools  
library(readr) # Read Rectangular Text Data  
library(rpart)  
library(rattle)  
library(scales) # Scale Functions for Visualization  
library(shiny) # Web Application Framework for R  
library(skimr) # Compact and Flexible Summaries of Data  
library(stringr) # Simple, Consistent Wrappers for Common String Operations  
library(tibble) # Simple Data Frames  
library(tidyr) # Easily Tidy Data with ‘spread()’ and ‘gather()’ Functions  
library(tidyverse) # Install and Load ‘Tidyverse’ Packages  
library(usmap)  
library(rworldmap)  
library(ggmap)  
library(maps)  
library(mapdata)  
library(cowplot)  
library("googleway")  
library("ggspatial")   
library("rnaturalearth")   
library("rnaturalearthdata")  
theme\_set(theme\_bw())  
library("sf")  
register\_google(key = "AIzaSyB3n0SkJTXxrbUkF2bkBJs7XdleT\_iJXsg", write = TRUE)

## Replacing old key (AIzaSyB3n0SkJTXxrbUkF2bkBJs7XdleT\_iJXsg) with new key in /Users/Paige/.Renviron

#### Overall histogram of counts per sample survey

SS\_count %>% ggplot() +   
 geom\_histogram(aes(total))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



#### Create new data frame that just lists each site once

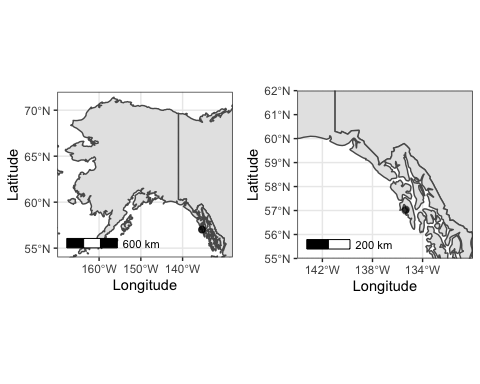
unique\_sites <- subset(raw\_SS\_count, !duplicated(raw\_SS\_count$site\_code))   
sites\_coordinates <- unique\_sites[c(2,5:6,15,22)]  
sites\_coordinates\_contig <- filter(sites\_coordinates, sites\_coordinates$state\_province != "Alaska")   
  
CA\_coordinates <- filter(sites\_coordinates, sites\_coordinates$state\_province == "California")  
OR\_coordinates <- filter(sites\_coordinates, sites\_coordinates$state\_province == "Oregon")  
WA\_coordinates <- filter(sites\_coordinates, sites\_coordinates$state\_province == "Washington")  
AK\_coordinates <- filter(sites\_coordinates, sites\_coordinates$state\_province == "Alaska")  
  
CA\_coordinates <- as.data.frame(CA\_coordinates)  
OR\_coordinates <- as.data.frame(OR\_coordinates)  
WA\_coordinates <- as.data.frame(WA\_coordinates)  
AK\_coordinates <- as.data.frame(AK\_coordinates)

#### Plot sample sites

After messing around with several different mapping tools in R, I decided that trying to display the whole range (CA to AK) on one map would be really hard, and the Alaska sites are really just at one specific location, so I’m going to split Alaska off from the three states that are part of the contiguous US #### Plot AK sites

world <- ne\_countries(scale = "medium", returnclass = "sf")  
AK\_zoom1 <- ggplot(data = world) +  
 geom\_sf() +  
 geom\_point(data= sites\_coordinates,aes(x=sites\_coordinates$longitude, y=sites\_coordinates$latitude,), size = 2, alpha = 0.4) +  
 annotation\_scale(location = "bl", width\_hint = 0.3) +  
 coord\_sf(xlim = c(-170, -128), ylim = c(54, 72), expand = FALSE) +   
 xlab("Longitude") +   
 ylab("Latitude") +   
 scale\_x\_continuous(breaks = c(-160, -150, -140))  
   
  
AK\_zoom2 <- ggplot(data = world) +  
 geom\_sf() +  
 geom\_point(data= sites\_coordinates,aes(x=sites\_coordinates$longitude, y=sites\_coordinates$latitude,), size = 2, alpha = 0.4) +  
 annotation\_scale(location = "bl", width\_hint = 0.3) +  
 coord\_sf(xlim = c(-144, -130), ylim = c(55, 62), expand = FALSE) +   
 xlab("Longitude") +   
 ylab("Latitude") +   
 scale\_x\_continuous(breaks = c(-142, -138, -134))  
   
grid.arrange(AK\_zoom1, AK\_zoom2, ncol = 2)

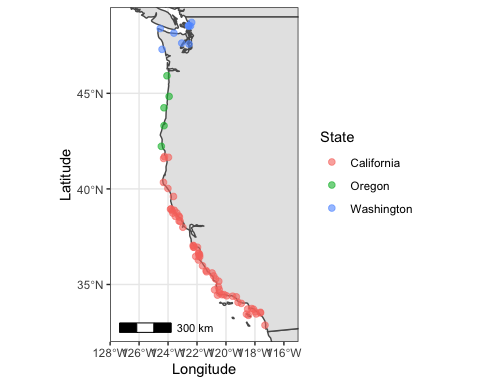
## Scale on map varies by more than 10%, scale bar may be inaccurate  
## Scale on map varies by more than 10%, scale bar may be inaccurate



#### Plot WA, OR, and CA sites

State <- sites\_coordinates\_contig$state\_province  
ggplot(data = world) +  
 geom\_sf() +  
 geom\_jitter(data= sites\_coordinates\_contig,aes(x=sites\_coordinates\_contig$longitude, y=sites\_coordinates\_contig$latitude,col=State), size = 2, alpha = 0.6, width = 0.2) +  
 annotation\_scale(location = "bl", width\_hint = 0.3) +  
 coord\_sf(xlim = c(-128, -115), ylim = c(32, 49.5), expand = FALSE) +   
 xlab("Longitude") +   
 ylab("Latitude")

## Scale on map varies by more than 10%, scale bar may be inaccurate



### Exploring Count Data by State and Year Surveyed

#### What does the structure of the data look like by state?

#### CA

CA <- SS\_count %>% select(marine\_site\_name, marine\_common\_year, size\_bin, total, state\_province) %>%  
 filter(state\_province == "California")  
summary(CA)

## marine\_site\_name marine\_common\_year size\_bin total   
## Length:10548 Min. :2000 Min. : 5.00 Min. : 1.000   
## Class :character 1st Qu.:2005 1st Qu.: 50.00 1st Qu.: 2.000   
## Mode :character Median :2009 Median : 80.00 Median : 5.000   
## Mean :2009 Mean : 81.92 Mean : 9.607   
## 3rd Qu.:2013 3rd Qu.:110.00 3rd Qu.: 11.000   
## Max. :2018 Max. :230.00 Max. :360.000   
## state\_province   
## Length:10548   
## Class :character   
## Mode :character   
##   
##   
##

#### OR

OR <- SS\_count %>% select(marine\_site\_name, marine\_common\_year, size\_bin, total, state\_province) %>%  
 filter(state\_province == "Oregon")  
summary(OR)

## marine\_site\_name marine\_common\_year size\_bin total   
## Length:1386 Min. :2000 Min. : 5.00 Min. : 1.00   
## Class :character 1st Qu.:2005 1st Qu.: 50.00 1st Qu.: 2.00   
## Mode :character Median :2010 Median : 80.00 Median : 7.00   
## Mean :2010 Mean : 87.21 Mean : 15.28   
## 3rd Qu.:2014 3rd Qu.:120.00 3rd Qu.: 18.00   
## Max. :2018 Max. :230.00 Max. :212.00   
## state\_province   
## Length:1386   
## Class :character   
## Mode :character   
##   
##   
##

#### WA

WA <- SS\_count %>% select(marine\_site\_name, marine\_common\_year, size\_bin, total, state\_province) %>%  
 filter(state\_province == "Washington")  
summary(WA)

## marine\_site\_name marine\_common\_year size\_bin total   
## Length:865 Min. :2008 Min. : 5.00 Min. : 1.000   
## Class :character 1st Qu.:2011 1st Qu.: 50.00 1st Qu.: 1.000   
## Mode :character Median :2014 Median : 90.00 Median : 2.000   
## Mean :2014 Mean : 89.26 Mean : 6.776   
## 3rd Qu.:2016 3rd Qu.:120.00 3rd Qu.: 5.000   
## Max. :2018 Max. :320.00 Max. :133.000   
## state\_province   
## Length:865   
## Class :character   
## Mode :character   
##   
##   
##

#### AK

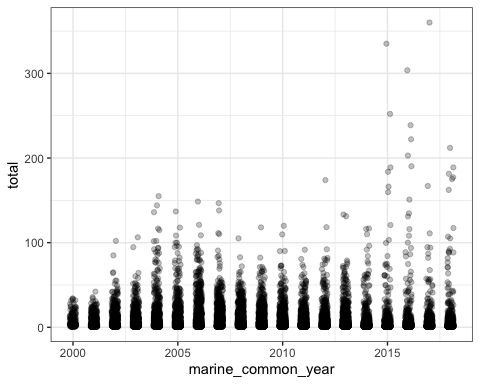
AK <- SS\_count %>% select(marine\_site\_name, marine\_common\_year, size\_bin, total, state\_province) %>%  
 filter(state\_province == "Alaska")  
summary(AK)

## marine\_site\_name marine\_common\_year size\_bin total   
## Length:366 Min. :2011 Min. : 5.00 Min. : 1.00   
## Class :character 1st Qu.:2013 1st Qu.: 40.00 1st Qu.: 1.00   
## Mode :character Median :2015 Median : 60.00 Median : 3.00   
## Mean :2015 Mean : 63.51 Mean : 6.03   
## 3rd Qu.:2017 3rd Qu.: 87.50 3rd Qu.: 7.00   
## Max. :2018 Max. :190.00 Max. :117.00   
## state\_province   
## Length:366   
## Class :character   
## Mode :character   
##   
##   
##

[INSERT TABLE SUMMARIZING THIS INFORMATION]

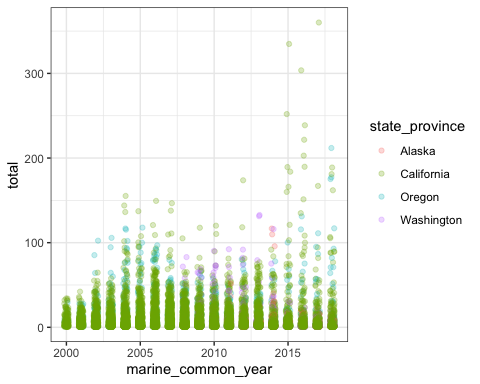
#### What does sea star abundance over time look like?

SS\_count %>% ggplot(aes(marine\_common\_year, total)) + geom\_jitter(width = 0.15, alpha = 0.25)



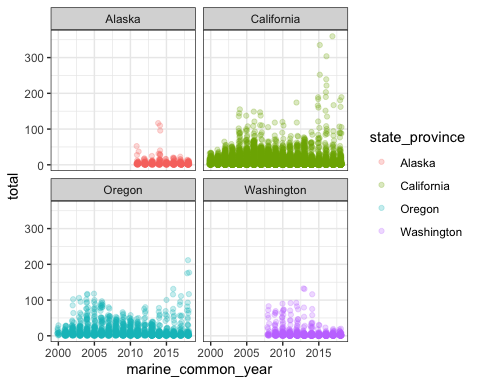
#### Lets color the points by state

SS\_count %>% ggplot(aes(marine\_common\_year, total, col=state\_province)) + geom\_jitter(width = 0.15, alpha = 0.25)

 Woah, California is dominating this visualization.

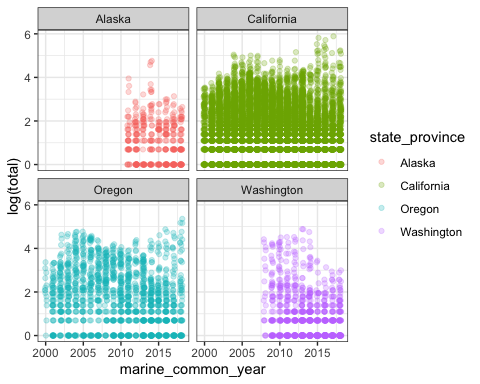
#### Lets break that up into 4 parts:

SS\_count %>%   
 ggplot(aes(marine\_common\_year, total, col=state\_province)) + geom\_jitter(width = 0.15, alpha = 0.25) +  
 facet\_wrap(~state\_province)

 Washington and Alaska clearly weren’t surveyed for all years during which California and Oregon were.

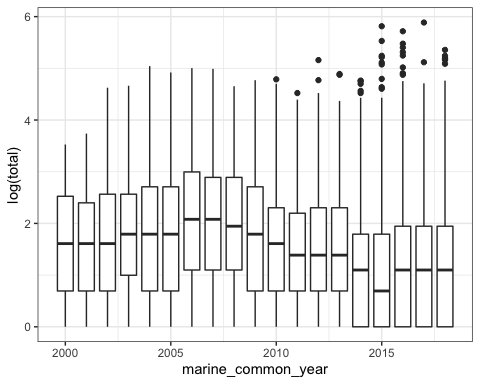
#### What do these counts look like if we display them as log()?

SS\_count %>%   
 ggplot(aes(marine\_common\_year, log(total), col=state\_province)) + geom\_jitter(width = 0.15, alpha = 0.25) +  
 facet\_wrap(~state\_province)



#### Abundance over time

SS\_count %>% ggplot(aes(marine\_common\_year, log(total), group = marine\_common\_year)) + geom\_boxplot()



#### So far I’ve been looking at this across all 3 species of sea stars surveyed. What does the last graph look like for each species, seperately?

#### Summary statistics for each species over time

PISOCH\_only <- filter(SS\_count, species\_code == "PISOCH")  
KATTUN\_only <- filter(SS\_count, species\_code == "KATTUN")  
EVATRO\_only <- filter(SS\_count, species\_code == "EVATRO")  
skim(PISOCH\_only)

## Skim summary statistics  
## n obs: 11416   
## n variables: 24   
##   
## ── Variable type:character ─────────────────────────────────────────────────────────  
## variable missing complete n min max empty n\_unique  
## bioregion 0 11416 11416 13 33 0 6  
## georegion 0 11416 11416 2 24 0 9  
## group\_code 0 11416 11416 2 6 0 9  
## island 0 11416 11416 3 9 0 7  
## last\_updated 0 11416 11416 19 19 0 1  
## marine\_season\_code 0 11416 11416 4 4 0 57  
## marine\_site\_name 0 11416 11416 5 42 0 77  
## method\_code 0 11416 11416 2 4 0 4  
## mpa\_designation 0 11416 11416 2 9 0 5  
## mpa\_region 0 11416 11416 2 19 0 6  
## season\_name 0 11416 11416 4 6 0 4  
## site\_code 0 11416 11416 3 4 0 77  
## species\_code 0 11416 11416 6 6 0 1  
## state\_province 0 11416 11416 6 10 0 4  
## target\_assemblage 0 11416 11416 8 8 0 1  
##   
## ── Variable type:integer ───────────────────────────────────────────────────────────  
## variable missing complete n mean sd p0 p25  
## marine\_common\_season 0 11416 11416 114.4 19.14 78 99  
## marine\_common\_year 0 11416 11416 2009.09 4.79 2000 2005  
## marine\_sort\_order 0 11416 11416 5682.02 1048.93 1720 5050  
## season\_sequence 0 11416 11416 2.03 0.84 1 1  
## size\_bin 0 11416 11416 86.83 42.67 5 50  
## size\_sort\_order 0 11416 11416 9.68 4.27 1 6  
## total 0 11416 11416 10.85 17.41 1 2  
## p50 p75 p100 hist  
## 114 130 152 ▅▅▆▇▆▆▅▃  
## 2009 2013 2018 ▃▅▅▇▆▆▃▅  
## 6150 6390 7650 ▁▁▁▂▃▂▇▁  
## 2 3 4 ▇▁▆▁▁▇▁▁  
## 90 120 320 ▅▇▇▅▁▁▁▁  
## 10 13 33 ▅▇▇▅▁▁▁▁  
## 5 13 360 ▇▁▁▁▁▁▁▁  
##   
## ── Variable type:numeric ───────────────────────────────────────────────────────────  
## variable missing complete n mean sd p0 p25 p50  
## latitude 0 11416 11416 38.13 4.82 32.87 34.55 36.51  
## longitude 0 11416 11416 -121.94 2.57 -135.38 -123.51 -121.91  
## p75 p100 hist  
## 40.34 57.05 ▇▅▂▂▁▁▁▁  
## -120.61 -117.25 ▁▁▁▁▅▇▇▅

skim(KATTUN\_only)

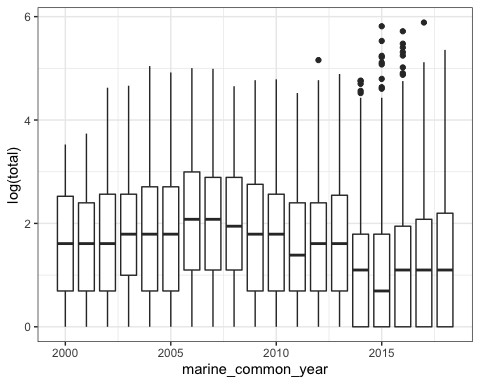
## Skim summary statistics  
## n obs: 1676   
## n variables: 24   
##   
## ── Variable type:character ─────────────────────────────────────────────────────────  
## variable missing complete n min max empty n\_unique  
## bioregion 0 1676 1676 13 33 0 4  
## georegion 0 1676 1676 2 16 0 7  
## group\_code 0 1676 1676 2 6 0 6  
## island 0 1676 1676 3 9 0 5  
## last\_updated 0 1676 1676 19 19 0 1  
## marine\_season\_code 0 1676 1676 4 4 0 27  
## marine\_site\_name 0 1676 1676 5 42 0 46  
## method\_code 0 1676 1676 2 2 0 1  
## mpa\_designation 0 1676 1676 2 9 0 5  
## mpa\_region 0 1676 1676 2 19 0 5  
## season\_name 0 1676 1676 4 6 0 3  
## site\_code 0 1676 1676 3 4 0 46  
## species\_code 0 1676 1676 6 6 0 1  
## state\_province 0 1676 1676 6 10 0 4  
## target\_assemblage 0 1676 1676 8 8 0 1  
##   
## ── Variable type:integer ───────────────────────────────────────────────────────────  
## variable missing complete n mean sd p0 p25 p50  
## marine\_common\_season 0 1676 1676 132.29 10.05 114 123 133  
## marine\_common\_year 0 1676 1676 2013.57 2.54 2009 2011 2014  
## marine\_sort\_order 0 1676 1676 4911.32 1220.64 1720 4400 5050  
## season\_sequence 0 1676 1676 2.01 0.62 1 2 2  
## size\_bin 0 1676 1676 53.79 23.85 5 35 50  
## size\_sort\_order 0 1676 1676 8.95 3.05 1 7 9  
## total 0 1676 1676 3.9 3.26 1 1 3  
## p75 p100 hist  
## 141 150 ▆▇▅▆▇▇▅▆  
## 2016 2018 ▇▇▅▆▇▇▆▇  
## 6130 6310 ▂▁▁▁▂▇▁▆  
## 2 3 ▂▁▁▇▁▁▁▂  
## 70 130 ▃▃▇▅▇▂▁▁  
## 11 17 ▂▂▃▇▇▅▁▁  
## 5 19 ▇▂▂▁▁▁▁▁  
##   
## ── Variable type:numeric ───────────────────────────────────────────────────────────  
## variable missing complete n mean sd p0 p25 p50  
## latitude 0 1676 1676 41.73 5.77 35.45 36.56 40.34  
## longitude 0 1676 1676 -124.07 3.33 -135.38 -124.14 -123.73  
## p75 p100 hist  
## 44.24 57.05 ▇▅▆▂▃▁▁▂  
## -121.95 -120.95 ▁▁▁▁▁▁▇▆

skim(EVATRO\_only)

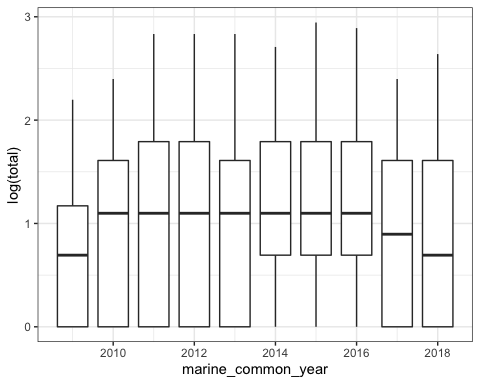
## Skim summary statistics  
## n obs: 73   
## n variables: 24   
##   
## ── Variable type:character ─────────────────────────────────────────────────────────  
## variable missing complete n min max empty n\_unique  
## bioregion 0 73 73 13 26 0 2  
## georegion 0 73 73 2 13 0 2  
## group\_code 0 73 73 2 6 0 5  
## island 0 73 73 3 9 0 6  
## last\_updated 0 73 73 19 19 0 1  
## marine\_season\_code 0 73 73 4 4 0 10  
## marine\_site\_name 0 73 73 9 42 0 10  
## method\_code 0 73 73 2 2 0 1  
## mpa\_designation 0 73 73 4 4 0 1  
## mpa\_region 0 73 73 4 4 0 1  
## season\_name 0 73 73 4 6 0 3  
## site\_code 0 73 73 3 4 0 10  
## species\_code 0 73 73 6 6 0 1  
## state\_province 0 73 73 6 10 0 2  
## target\_assemblage 0 73 73 8 8 0 1  
##   
## ── Variable type:integer ───────────────────────────────────────────────────────────  
## variable missing complete n mean sd p0 p25 p50  
## marine\_common\_season 0 73 73 140.75 9.19 114 134 142  
## marine\_common\_year 0 73 73 2015.7 2.28 2009 2014 2016  
## marine\_sort\_order 0 73 73 2496.03 833.39 1720 1720 1740  
## season\_sequence 0 73 73 1.96 0.26 1 2 2  
## size\_bin 0 73 73 54.11 28.41 5 30 50  
## size\_sort\_order 0 73 73 6.4 2.87 1 4 6  
## total 0 73 73 1.84 1.83 1 1 1  
## p75 p100 hist  
## 150 150 ▁▁▁▂▃▁▃▇  
## 2018 2018 ▁▁▁▂▃▁▃▇  
## 3290 3850 ▇▁▁▁▁▃▃▁  
## 2 3 ▁▁▁▇▁▁▁▁  
## 70 160 ▅▇▆▅▅▁▁▁  
## 8 17 ▅▇▆▅▅▁▁▁  
## 2 13 ▇▁▁▁▁▁▁▁  
##   
## ── Variable type:numeric ───────────────────────────────────────────────────────────  
## variable missing complete n mean sd p0 p25 p50 p75  
## latitude 0 73 73 52.86 4.5 47.58 48.52 56.99 57.05  
## longitude 0 73 73 -129.42 6.4 -135.38 -135.35 -135.32 -122.55  
## p100 hist  
## 57.05 ▇▁▁▁▁▁▁▇  
## -122.52 ▇▁▁▁▁▁▁▇

#### Species abundance over time

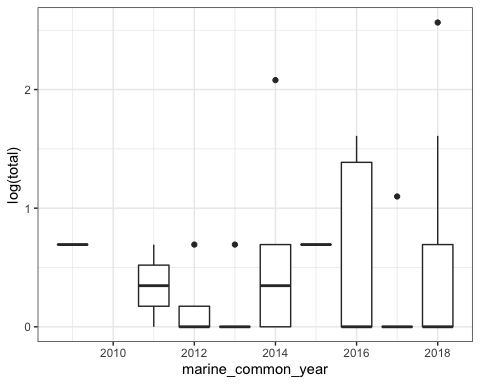
PISOCH\_only %>% ggplot(aes(marine\_common\_year, log(total), group = marine\_common\_year)) + geom\_boxplot()



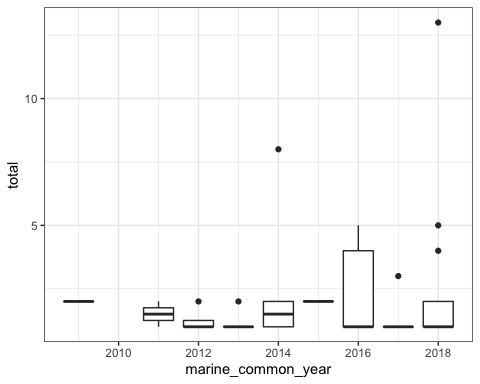
KATTUN\_only %>% ggplot(aes(marine\_common\_year, log(total), group = marine\_common\_year)) + geom\_boxplot()



EVATRO\_only %>% ggplot(aes(marine\_common\_year, log(total), group = marine\_common\_year)) + geom\_boxplot()

 That last one looked like the log() function might not be applicable. Let’s see what it looks like without that:

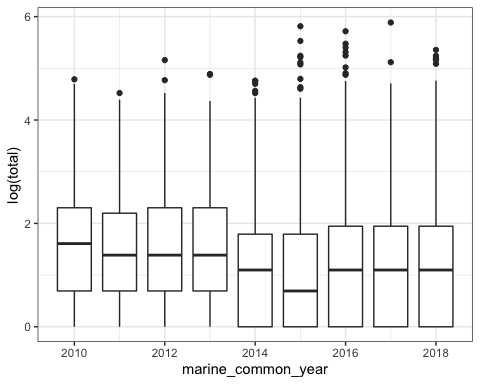
EVATRO\_only %>% ggplot(aes(marine\_common\_year, total, group = marine\_common\_year)) + geom\_boxplot()



#### Characterizing sea star abundance in the last decade

The most recent and catastrophic sea star wasting disease outbreak hit the west coast of North America around 2013. How did sea star abundance look before the epidemic, how did it change during the epidemic, and have the populations rebounded?

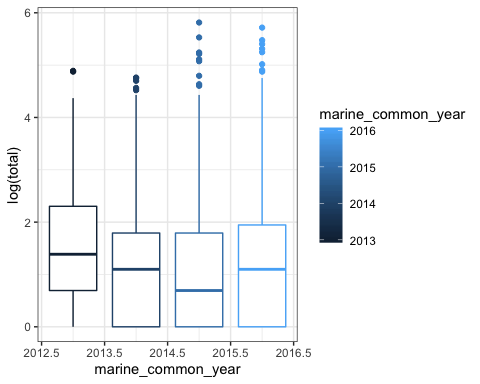
last\_decade <- filter(SS\_count, marine\_common\_year >= "2010")  
last\_decade %>% ggplot(aes(marine\_common\_year, log(total), group = marine\_common\_year)) + geom\_boxplot()

 It looks like populations leveled back out at 2016 and remained relatively unchanged in 2017 and 2018.

#### Narrowing the time frame under analysis

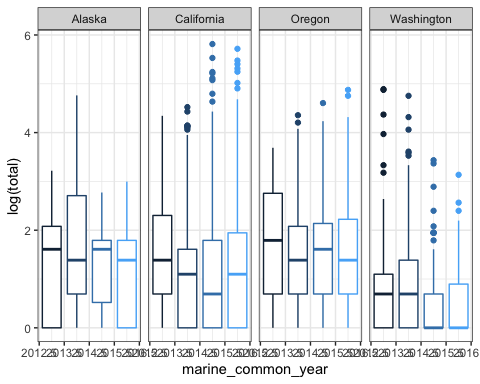
It looks like theres a detectable decrease between 2013 and 2014, and that populations leveled back out at 2016 and remained relatively unchanged in 2017 and 2018. Lets narrow the time frame to only include sampling between 2013 and 2016.

last\_decade$marine\_common\_year <- as.factor(last\_decade$marine\_common\_year)  
SSWD\_timeframe <- filter(SS\_count, marine\_common\_year >= "2013" & marine\_common\_year <= "2016")  
SSWD\_timeframe %>% ggplot(aes(marine\_common\_year, log(total), group = marine\_common\_year, col=marine\_common\_year)) + geom\_boxplot()



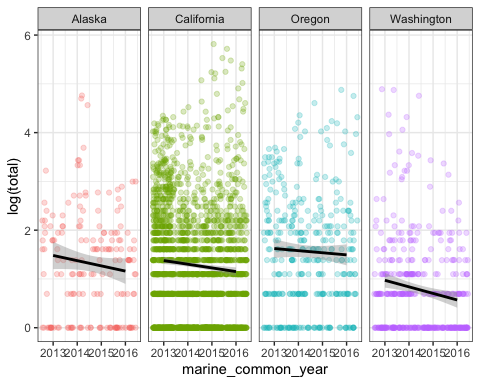
#### By state

SSWD\_timeframe %>% ggplot(aes(marine\_common\_year, log(total), group = marine\_common\_year, col=marine\_common\_year)) + geom\_boxplot() + facet\_wrap(~state\_province, 1) + theme(legend.position = "none")



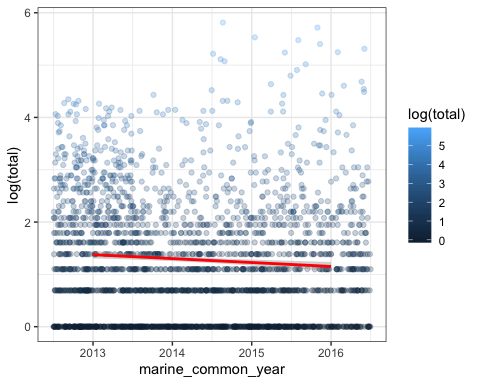
#### Fitting a linear model

SSWD\_timeframe %>% ggplot(aes(marine\_common\_year, log(total), col=state\_province)) + geom\_jitter(width = 0.45, alpha = 0.25) +  
 facet\_wrap(~state\_province, 1) + geom\_smooth(method = "lm", col="black") + theme(legend.position = "none")



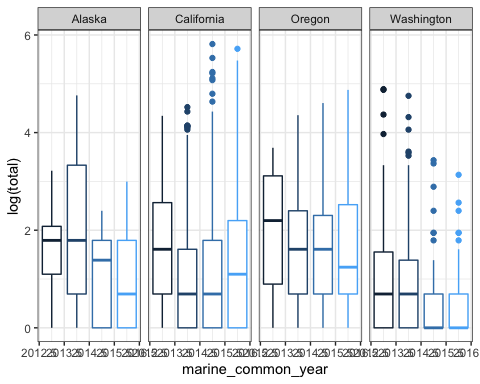
#### Just California

SSWD\_timeframe\_CA <- filter(SSWD\_timeframe, state\_province == "California")  
SSWD\_timeframe\_CA %>% ggplot(aes(marine\_common\_year, log(total), col=log(total))) + geom\_jitter(width = 0.5, alpha = 0.25) + geom\_smooth(method = "lm", col="red")

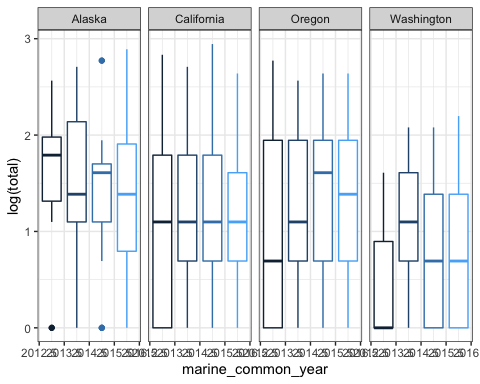


#### Doing the same as above, but separating by species, with a primary focus on Pisaster ochraceus (the organism I’m studing in my research)

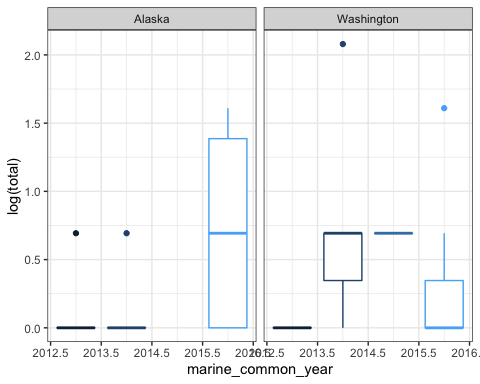
SSWD\_timeframe\_PIS <- filter(PISOCH\_only, marine\_common\_year >= "2013" & marine\_common\_year <= "2016")  
SSWD\_timeframe\_PIS %>% ggplot(aes(marine\_common\_year, log(total), group = marine\_common\_year, col=marine\_common\_year)) + geom\_boxplot() + facet\_wrap(~state\_province, 1) + theme(legend.position = "none")



SSWD\_timeframe\_KAT <- filter(KATTUN\_only, marine\_common\_year >= "2013" & marine\_common\_year <= "2016")  
SSWD\_timeframe\_KAT %>% ggplot(aes(marine\_common\_year, log(total), group = marine\_common\_year, col=marine\_common\_year)) + geom\_boxplot() + facet\_wrap(~state\_province, 1) + theme(legend.position = "none")



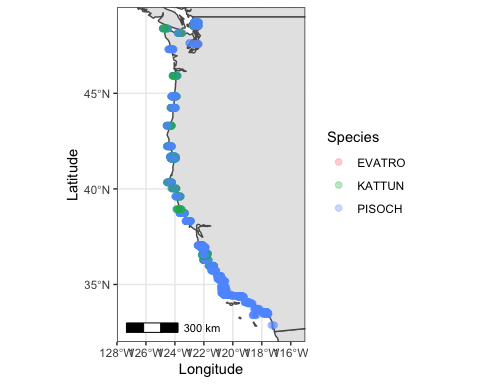
SSWD\_timeframe\_EVA <- filter(EVATRO\_only, marine\_common\_year >= "2013" & marine\_common\_year <= "2016")  
SSWD\_timeframe\_EVA %>% ggplot(aes(marine\_common\_year, log(total), group = marine\_common\_year, col=marine\_common\_year)) + geom\_boxplot() + facet\_wrap(~state\_province, 1) + theme(legend.position = "none")

 OH! I didn’t catch that the reason why the EVA species had such few survey sightings was because its range appears to only be in Alaska. Speaking of which….

#### I should go back and plot each survey entry and color by species observed to get an idea of the ranges of each.

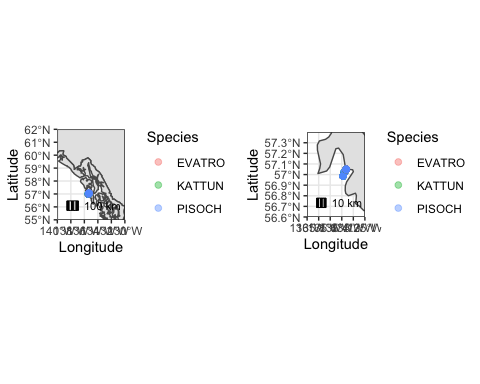
Species <- SSWD\_timeframe$species\_code  
ggplot(data = world) +  
 geom\_sf() +  
 geom\_jitter(data= SSWD\_timeframe,aes(x=SSWD\_timeframe$longitude, y=SSWD\_timeframe$latitude,col=Species), size = 2, alpha = 0.3, width = 0.2) +  
 annotation\_scale(location = "bl", width\_hint = 0.3) +  
 coord\_sf(xlim = c(-128, -115), ylim = c(32, 49.5), expand = FALSE) +   
 xlab("Longitude") +   
 ylab("Latitude")

## Scale on map varies by more than 10%, scale bar may be inaccurate



AK\_zoom1\_SSWD <- ggplot(data = world) +  
 geom\_sf() +  
 geom\_point(data= SSWD\_timeframe,aes(x=SSWD\_timeframe$longitude, y=SSWD\_timeframe$latitude,col=Species,), size = 2, alpha = 0.4) +  
 annotation\_scale(location = "bl", width\_hint = 0.3) +  
 coord\_sf(xlim = c(-140, -130), ylim = c(55, 62), expand = FALSE) +   
 xlab("Longitude") +   
 ylab("Latitude")   
   
  
AK\_zoom2\_SSWD <- ggplot(data = world) +  
 geom\_sf() +  
 geom\_point(data= SSWD\_timeframe,aes(x=SSWD\_timeframe$longitude, y=SSWD\_timeframe$latitude,col=Species,), size = 2, alpha = 0.4) +  
 annotation\_scale(location = "bl", width\_hint = 0.3) +  
 coord\_sf(xlim = c(-136, -135), ylim = c(56.6, 57.4), expand = FALSE) +   
 xlab("Longitude") +   
 ylab("Latitude")   
   
grid.arrange(AK\_zoom1\_SSWD, AK\_zoom2\_SSWD, ncol = 2)

## Scale on map varies by more than 10%, scale bar may be inaccurate



## Other components (unfinished)

### Methods and Results

*In most research papers, results and methods are separate. You can combine them here if you find it easier. You are also welcome to structure things such that those are separate sections.*

#### Data aquisition

*As applicable, explain where and how you got the data. If you directly import the data from an online source, you can combine this section with the next.*

#### Data import and cleaning

*Write code that reads in the file and cleans it so it’s ready for analysis. Since this will be fairly long code for most datasets, it might be a good idea to have it in one or several R scripts. If that is the case, explain here briefly what each file does. The files themselves should be commented well so everyone can follow along.*

#### Univariate analysis

*Use a combination of text/tables/figures to explore and describe your data. You should produce plots or tables or other summary quantities for most of your variables. You definitely need to do it for the important variables, i.e. if you have main exposure or outcome variables, those need to be explored. Depending on the total number of variables in your dataset, explore all or some of the others.*

#### Bivariate analysis

*Create plots or tables and compute simple statistics (e.g. t-tests, simple regression model with 1 predictor, etc.) to look for associations between your outcome(s) and each individual predictor variable*

#### Full analysis

*Use one or several suitable statistical/machine learning methods to analyze your data and to produce meaningful figures, tables, etc. This might again be code that is best placed in one or several separate R scripts that need to be well documented. You can then load the results produced by this code*

### Discussion

#### Summary and Interpretation

*Summarize what you did, what you found and what it means.*

#### Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

#### Conclusions

*What are the main take-home messages?*

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

#### References