Citizen\_Science\_Efficacy\_Analysis

William Norfolk

11/19/2019

This script compares the MarineLab citizen science water quality database to the Florida Keys National Marine Scanctuary (FKNMS) water quality monitoring database to determine how similar/dissimilar each of these data sets are.

It should be noted that the sampling sites for FKNMS and MarineLab are not the same due to differences in the needs of the programs. Comparisions will only be made with sites which are geographic (or very close) replicates of each other.

First begin by loading all of the libraries needed for data cleaning and visualization.

library(readxl)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidyverse)

## -- Attaching packages --------------------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.2.1 v readr 1.3.1  
## v tibble 2.1.3 v purrr 0.3.3  
## v tidyr 1.0.0 v stringr 1.4.0  
## v ggplot2 3.2.1 v forcats 0.4.0

## -- Conflicts ------------------------------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(forcats)  
library(ggthemes)  
library(knitr)  
library(naniar)  
library(broom)  
library(gridExtra)

##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

library(zoo)

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

library(ggpubr)

## Loading required package: magrittr

##   
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':  
##   
## set\_names

## The following object is masked from 'package:tidyr':  
##   
## extract

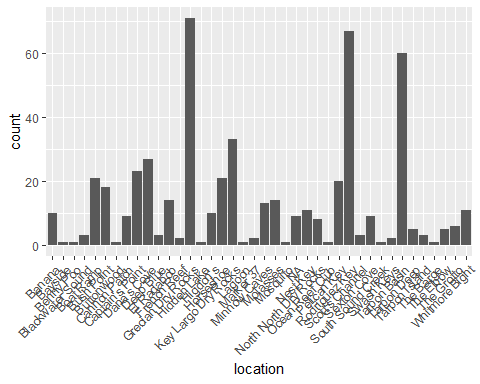
Next load the MarineLab citizen science data.

WQ\_clean\_data <- readRDS("../../data/processed\_data/processeddata.rds")  
  
glimpse(WQ\_clean\_data)

## Observations: 522  
## Variables: 15  
## $ Month <chr> "01", "02", "02", "02", "02", "02", "02", "02...  
## $ Day <chr> "08", "08", "08", "08", "09", "09", "09", "12...  
## $ Year <chr> "16", "16", "16", "16", "16", "16", "16", "16...  
## $ military\_time <dbl> 1415, 1515, 1550, 1555, 1001, 1015, 1022, 103...  
## $ location <chr> "Boat Ramp", "Grecian Dry Rocks", "Grecian Dr...  
## $ instructor\_name <chr> "Katy, Sarah, Driver", "Chelsea", "Katy, Tomm...  
## $ group\_name <chr> "NA", "McLean High School", "McLean High Scho...  
## $ ph <dbl> 8.0, 8.4, 8.2, 8.4, 8.0, 8.0, 8.0, 8.0, 8.4, ...  
## $ ammonia <dbl> 0.00, 0.00, 0.00, 0.00, 0.25, 0.00, 0.00, 0.0...  
## $ dissolved\_oxygen <dbl> 5.0, 4.0, 4.0, 6.0, 8.0, 4.0, 5.0, 6.0, 6.0, ...  
## $ water\_temp <dbl> NA, 23.5, 21.0, 36.0, 18.0, 18.0, 18.0, 18.3,...  
## $ salinity <dbl> 36, 40, 44, 35, 33, 30, 33, 35, 40, 30, 35, 2...  
## $ equipment <chr> "kit", "kit", "kit", "kit", "kit", "kit", "ki...  
## $ island\_side <chr> "ocean", "ocean", "ocean", "ocean", NA, "bay"...  
## $ site\_type <chr> "Seagrass/Mangrove", "Coral Reef", "Coral Ree...

Before we begin to work with the new dataset we need to make a reference for the citizen science data so select comparable sites. There may be some site overlap however, this overlap is not very meaningful if the n for that site is very small for the citizen science data. FKNMS data is on a standardized sampling schedule so it is unlikely this will be an issue with that data set. We will make a simple bar chart of the site counts for reference.

site\_view <- ggplot(WQ\_clean\_data, aes(x = location)) + geom\_bar(stat = "count") + theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
  
site\_view

 Ideally, we can select all of the sites where n > 20 however, some of these locations will likely not be present in the FKNMS data or may be too specific for there sampling needs. For certian we will use our three representitive sites Grecian Dry Rocks, Rodriguez Key, and Tarpon Basin; other contenders for comparison are Boat Ramp, Captian’s Point, Dana’s Cove, Horseshoe, Key Largo Dry Rocks, and Pelican Key. We will explore this data in the subsequent steps.

Next we will load the raw FKNMS water quality data. This data sheet can be found in the raw\_data folder within the code folder of this project repository.

fknms\_data\_raw <- readxl::read\_excel("../../data/raw\_data/FKNMS\_WQ\_raw.xlsx")

## New names:  
## \* YEAR -> YEAR...8  
## \* YEAR -> YEAR...19  
## \* `` -> ...65  
## \* `` -> ...66  
## \* `` -> ...67  
## \* ... and 32 more problems

glimpse(fknms\_data\_raw)

## Observations: 13,427  
## Variables: 99  
## $ `Survey No.` <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15...  
## $ DATE <chr> "34787", "34905", "35038", "35137", "35285", "354...  
## $ TIME <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, "...  
## $ STA <dbl> 200, 200, 200, 200, 200, 200, 200, 200, 200, 200,...  
## $ SITE <chr> "Fowey Rocks", "Fowey Rocks", "Fowey Rocks", "Fow...  
## $ LATDEC <dbl> 25.59, 25.59, 25.59, 25.59, 25.59, 25.59, 25.59, ...  
## $ LONDEC <dbl> -80.1, -80.1, -80.1, -80.1, -80.1, -80.1, -80.1, ...  
## $ YEAR...8 <dbl> 1995.237, 1995.560, 1995.924, 1996.195, 1996.600,...  
## $ DEPTH <dbl> 5.5, 5.5, 5.5, 5.5, 5.5, 5.5, 5.5, 5.5, 5.5, 5.5,...  
## $ ZONE <chr> "KR", "KR", "KR", "KR", "KR", "KR", "KR", "KR", "...  
## $ SEG <dbl> 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9...  
## $ CLUST <dbl> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3...  
## $ ZONE5 <chr> "REEF", "REEF", "REEF", "REEF", "REEF", "REEF", "...  
## $ ZONE6 <chr> "REEF", "REEF", "REEF", "REEF", "REEF", "REEF", "...  
## $ ZSEG <chr> "9-Reef", "9-Reef", "9-Reef", "9-Reef", "9-Reef",...  
## $ TRAN <chr> "BISC", "BISC", "BISC", "BISC", "BISC", "BISC", "...  
## $ NMILE <dbl> 3.5, 3.5, 3.5, 3.5, 3.5, 3.5, 3.5, 3.5, 3.5, 3.5,...  
## $ QUART <dbl> 2, 3, 4, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3...  
## $ YEAR...19 <dbl> 1995, 1995, 1995, 1996, 1996, 1996, 1997, 1997, 1...  
## $ `YR-QT` <chr> "1995-02", "1995-03", "1995-04", "1996-02", "1996...  
## $ `NOX-S` <dbl> 0.07750000, 0.04500000, 0.12250000, 0.01750000, 0...  
## $ `NOX-B` <dbl> NA, NA, NA, NA, 0.0200000, 0.0400000, 0.0675000, ...  
## $ `NO3-S` <dbl> 0.0200000000, NA, 0.0975000000, NA, NA, NA, 0.045...  
## $ `NO3-B` <dbl> NA, NA, NA, NA, NA, NA, NA, NA, 0.3600000, 0.0725...  
## $ `NO2-S` <dbl> 0.060000000, 0.045000000, 0.025000000, 0.01750000...  
## $ `NO2-B` <dbl> NA, NA, NA, NA, 0.0200000, 0.0400000, 0.0675000, ...  
## $ `NH4-S` <dbl> 0.2750000, 0.2750000, 0.2025000, 0.1350000, 0.122...  
## $ `NH4-B` <dbl> NA, NA, NA, NA, 0.0875000, 0.1825000, 0.2475000, ...  
## $ `TN-S` <dbl> 15.376079, 13.629270, 9.591759, 8.267623, 7.73550...  
## $ `TN-B` <dbl> NA, NA, NA, NA, 7.047500, 13.374286, 9.620859, 8....  
## $ `DIN-S` <dbl> 0.3525000, 0.3200000, 0.3250000, 0.1525000, 0.132...  
## $ `DIN-B` <dbl> NA, NA, NA, NA, 0.1075000, 0.2225000, 0.3150000, ...  
## $ `TON-S` <dbl> 15.023579, 13.309270, 9.266759, 8.115123, 7.60300...  
## $ `TON-B` <dbl> NA, NA, NA, NA, 6.940000, 13.151786, 9.305859, 8....  
## $ `TP-S` <dbl> 0.16250400, 0.17500431, 0.12250301, 0.10500258, 0...  
## $ `TP-B` <dbl> NA, NA, NA, NA, 0.13250326, 0.14500357, 0.1375033...  
## $ `SRP-S` <dbl> 0.012500308, 0.005000123, 0.002500062, NA, 0.0250...  
## $ `SRP-B` <dbl> NA, NA, NA, NA, 0.030000738, 0.017500431, NA, 0.0...  
## $ `APA-S` <dbl> 0.05512535, 0.07380222, 0.01104704, 0.03809723, 0...  
## $ `APA-B` <dbl> NA, NA, NA, NA, 0.04030000, 0.02220000, 0.0300000...  
## $ `CHLA-S` <dbl> 0.47805942, 0.06348215, 0.28286095, 0.24701600, 0...  
## $ `CHLA-B` <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ `TOC-S` <dbl> 313.3125, 227.0000, 135.4583, 129.5625, 158.6042,...  
## $ `TOC-B` <dbl> NA, NA, NA, NA, 165.2083, 189.7083, 234.9167, 267...  
## $ `SiO2-S` <dbl> NA, NA, NA, NA, 0.4572, 0.1498, 0.0675, 0.6825, 0...  
## $ `SiO2-B` <dbl> NA, NA, NA, NA, 0.457200000, 0.224700000, 0.06750...  
## $ `TURB-S` <dbl> 1.400, 0.600, 0.600, 0.300, 0.265, 0.150, 0.470, ...  
## $ `TURB-B` <dbl> NA, NA, NA, NA, 0.175, 0.310, 0.300, 0.100, 0.450...  
## $ `SAL-S` <dbl> 35.5000, 35.4000, 36.2000, 36.1000, 36.0000, 36.2...  
## $ `SAL-B` <dbl> 35.5000, 35.4000, 36.2000, 36.1000, NA, 36.2000, ...  
## $ `TEMP-S` <dbl> 23.6000, 30.3000, 25.6000, 23.3000, 31.0000, 24.8...  
## $ `TEMP-B` <dbl> 23.6000, 30.3000, 25.6000, 23.3000, NA, 24.3000, ...  
## $ `DO-S` <dbl> 7.00000, 6.40000, 6.70000, 6.80000, 6.25000, 5.90...  
## $ `DO-B` <dbl> 7.00000, 6.40000, 6.70000, 6.80000, NA, 6.10000, ...  
## $ Kd <chr> NA, NA, NA, NA, NA, "2.087E-2", NA, "0.245", "0.3...  
## $ pH <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ `TN:TP` <dbl> 94.622024, 77.881544, 78.300073, 78.739269, 73.67...  
## $ `N:P` <dbl> 28.200000, 64.000000, 130.000000, NA, 5.300000, 4...  
## $ `DIN:TP` <dbl> 2.1692310, 1.8285714, 2.6530612, 1.4523810, 1.261...  
## $ `%SAT-S` <dbl> 71.32471, 97.92023, 98.99452, 97.89006, 96.42675,...  
## $ `%SAT-B` <dbl> 71.37029, 97.92023, 98.99452, 97.89006, NA, 88.88...  
## $ `%Io` <dbl> NA, NA, NA, NA, NA, 86.40799325, NA, 15.92149664,...  
## $ DSIGT <dbl> NA, NA, NA, NA, NA, 0.17044700, 0.07463620, 0.308...  
## $ `Si:DIN` <dbl> NA, NA, NA, 0.000000000, 3.450566038, 0.495206612...  
## $ ...65 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...66 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...67 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...68 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...69 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...70 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...71 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...72 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...73 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...74 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...75 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...76 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...77 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...78 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...79 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...80 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...81 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...82 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...83 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...84 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...85 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...86 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...87 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...88 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...89 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...90 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...91 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...92 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...93 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...94 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...95 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...96 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...97 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...98 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ ...99 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...

This data set is in need to some serious cleaning, luckily to compare to our citizen science data we only need a small porition of these variables. We will remove variables then take a closer look at cleaning criteria.

Unfortunately it looks like the FKNMS data did not collect pH regularly so to avoid 1000+ NA values we will have to exclude pH from our comparison.

fknms\_reduce\_vars <- fknms\_data\_raw %>% dplyr::select(SITE, `NH4-S`, `TEMP-S`, `DO-S`, `SAL-S`)  
  
#We will rename these to match our predictors in the citizen science dataset.  
fknms\_rename <- fknms\_reduce\_vars %>% rename(location = SITE,   
 water\_temp = `TEMP-S`,  
 ammonia = `NH4-S`,  
 dissolved\_oxygen = `DO-S`,  
 salinity = `SAL-S`)  
  
unique(fknms\_rename$location)

## [1] "Fowey Rocks"   
## [2] "Sands Key"   
## [3] "Bowles Bank"   
## [4] "Triumph Reef"   
## [5] "Elliott Key"   
## [6] "Margo Fish Shoal"   
## [7] "Ajax Reef"   
## [8] "Old Rhodes Key"   
## [9] "Old Rhodes Key Channel"   
## [10] "Channel Key"   
## [11] "Old Rhodes Key Reef"   
## [12] "Pennikamp G27"   
## [13] "Turtle Harbor"   
## [14] "Turtle Reef"   
## [15] "Port Elizabeth"   
## [16] "Carysfort Channel"   
## [17] "Carysfort Reef"   
## [18] "Rattlesnake Key"   
## [19] "White Bank"   
## [20] "The Elbow"   
## [21] "Radabob Key"   
## [22] "Radabob Key Channel"   
## [23] "Dixie Shoal"   
## [24] "Mosquito Bank"   
## [25] "Molasses Reef Channel"   
## [26] "Molasses Reef"   
## [27] "Tavernier Harbor"   
## [28] "Triangles"   
## [29] "Conch Reef"   
## [30] "Plantation Point"   
## [31] "The Rocks"   
## [32] "Davis Reef"   
## [33] "Upper Matecumbe Key"   
## [34] "Upper MateCumbe Chnl"   
## [35] "Fish Haven"   
## [36] "Lower Matecumbe Key"   
## [37] "Alligator Shoal"   
## [38] "Alligator Reef"   
## [39] "Matecumbe Harbor"   
## [40] "Lower Matecumbe Chnl"   
## [41] "Matecumbe Offshore"   
## [42] "Long Key"   
## [43] "Long Key Channel"   
## [44] "Tennessee Reef"   
## [45] "Long Key Pass Inshore"   
## [46] "Long Key Pass Channel"   
## [47] "Long Key Pass Offshore"   
## [48] "Key Colony Beach"   
## [49] "Coffins Patch Channel"   
## [50] "Coffins Patch Offshore"   
## [51] "Seven Mile Bridge"   
## [52] "Seven Mile Br. Channel"   
## [53] "Sombrero Key"   
## [54] "Spanish Harbor Keys"   
## [55] "Bahia Honda Key"   
## [56] "Bahia Honda Channel"   
## [57] "Bahia Honda Offshore"   
## [58] "Long Beach"   
## [59] "Big Pine Channel"   
## [60] "Big Pine Shoal"   
## [61] "Newfound Harbor Keys"   
## [62] "American Shoal Channel"   
## [63] "Looe Key Channel"   
## [64] "Looe Key"   
## [65] "Aquarius"   
## [66] NA   
## [67] "Tarpon Creek"   
## [68] "American Shoal"   
## [69] "Saddlebunch Keys"   
## [70] "West Washerwoman"   
## [71] "Maryland Shoal"   
## [72] "Boca Chica Key"   
## [73] "Eastern Sambo"   
## [74] "Eastern Sambo Offshore"   
## [75] "Boca Chica Channel"   
## [76] "Boca Chica Mid"   
## [77] "Western Sambo"   
## [78] "Key West Cut A"   
## [79] "Western Head"   
## [80] "Main Ship Channel"   
## [81] "Eastern Dry Rocks"   
## [82] "Middle Ground"   
## [83] "Arsenic Bank"   
## [84] "Tripod Bank"   
## [85] "Channel Key Pass"   
## [86] "Toms Harbor Cut"   
## [87] "Bamboo Banks"   
## [88] "Bamboo Key"   
## [89] "Bluefish Bank"   
## [90] "Bullard Bank"   
## [91] "John Sawyer Bank"   
## [92] "Bethel Bank"   
## [93] "Red Bay Bank"   
## [94] "Bullfrog Banks"   
## [95] "W. Bahia Honda Key"   
## [96] "Coconut Key"   
## [97] "Harbor Key Bank"   
## [98] "Bogie Channel"   
## [99] "Little Pine Key"   
## [100] "Cutoe Key"   
## [101] "Content Passage"   
## [102] "Pine Channel"   
## [103] "Toptree Hammock Chan."   
## [104] "Cudjoe Key"   
## [105] "Johnson Key Channel"   
## [106] "Tarpon Belly Keys"   
## [107] "Kemp Channel"   
## [108] "Marvin Key Channel"   
## [109] "Snipe Keys"   
## [110] "Shark Key"   
## [111] "E. Harbor Key Channel"   
## [112] "Lower Harbor Keys"   
## [113] "Bluefish Channel"   
## [114] "Calda Channel"   
## [115] "Man of War Harbor"   
## [116] "Garrison Bight"   
## [117] "KW Northwest Channel"   
## [118] "N Boca Grande Channel"   
## [119] "Loggerhead Marker"   
## [120] "Loggerhead Channel"   
## [121] "Satan Shoal"   
## [122] "Ellis Rock"   
## [123] "SE Marquesas"   
## [124] "N Quicksands"   
## [125] "Marquesas Rock"   
## [126] "New Ground"   
## [127] "S Quicksands"   
## [128] "Half Moon Shoal"   
## [129] "Rebecca Shoal"   
## [130] "Garden Key"   
## [131] "Northwest Channel"   
## [132] "NE DTNP"   
## [133] "N DTNP"   
## [134] "Southeast Channel"   
## [135] "W DTNP"   
## [136] "Loggerhead Offshore"   
## [137] "Hospital Key"   
## [138] "Logerhead Inshore"   
## [139] "Grecian Rocks"   
## [140] "White Shoal"   
## [141] "Lake Largo Canal, Key Largo"   
## [142] "Calusa Park Marina, Key Largo"   
## [143] "Priv. Docks, Upper Matacumbe"   
## [144] "Blackwood Dr. boat ramp, Islamorada"   
## [145] "100th St. Canal, Marathon"   
## [146] "Hidden Harbor Beach, Marathon"   
## [147] "Newfound Harbor Channel, Little Torch"  
## [148] "Doctor's Arm, Big Pine Key"   
## [149] "Marriot Beachside, Key West"   
## [150] "Key West International Airport"

Now we can clean some of these. We are only interested in samples from sites that overlap with our citizen science data so we will select for those samples. Note the the FKNMS data is collected throughout the enterity of the Florida Keys so much of the sample sites in this dataset are from the middle to lower keys. The overlapping site in this data set are Dixie Shoal, Molasses Reef, Tarpon Creek, Grecian Rocks, and Molasses Reef Channel. Note there are a few additional overlapping sites however, they have a low number of observations (n<20) thus we will keep them out of our comparitive analysis.

fknms\_1 <- fknms\_rename[fknms\_rename$location == "Dixie Shoal", ]  
fknms\_5 <- fknms\_rename[fknms\_rename$location == "Molasses Reef", ]  
fknms\_6 <- fknms\_rename[fknms\_rename$location == "Tarpon Creek", ]  
fknms\_8 <- fknms\_rename[fknms\_rename$location == "Grecian Rocks", ]  
fknms\_9 <- fknms\_rename[fknms\_rename$location == "Molasses Reef Channel", ]  
  
fknms\_total <- rbind(fknms\_1, fknms\_5, fknms\_6, fknms\_8, fknms\_9)  
  
fknms\_na\_drop <- na.omit(fknms\_total)

Now the data is looking more compareable. The next step is to recode the FKNMS sites with the correcponding site names from the citizen science data (there are many synonyms for specific locations).

fknms\_fix\_names <- fknms\_na\_drop %>% mutate(location = recode(location,   
 "Dixie Shoal" = "Key Largo Dry Rocks",  
 "Molasses Reef" = "Molasses",  
 "Tarpon Creek" = "Tarpon Basin",  
 "Grecian Rocks" = "Grecian Dry Rocks",  
 "Molasses Reef Channel" = "Rodriguez Key"))

Next we need to select the corresponding citizen science data from these locaitons and bind them into a new dataframe.

cs\_1 <- WQ\_clean\_data[WQ\_clean\_data$location == "Grecian Dry Rocks", ]  
cs\_2 <- WQ\_clean\_data[WQ\_clean\_data$location == "Key Largo Dry Rocks", ]  
cs\_4 <- WQ\_clean\_data[WQ\_clean\_data$location == "Molasses", ]  
cs\_5 <- WQ\_clean\_data[WQ\_clean\_data$location == "Tarpon Basin", ]  
cs\_9 <- WQ\_clean\_data[WQ\_clean\_data$location == "Rodriguez Key", ]  
  
cs\_total <- rbind(cs\_1, cs\_2, cs\_4, cs\_5, cs\_9)

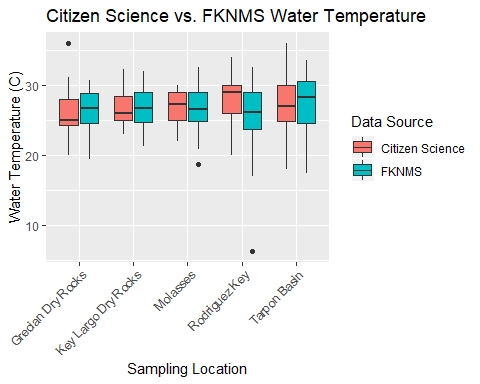
Next we need to drop the unused variables from the citizen science data and add a ned variable which we will call source to indicate where the data is coming from (Citizen Science or FKNMS). Then we can combine the two dataframes into one and create factor bins for the specific locations and sources.

cs\_reduce\_vars <- cs\_total %>% dplyr::select(location, water\_temp, salinity, dissolved\_oxygen, ammonia)  
  
cs\_add\_source <- cs\_reduce\_vars %>% mutate(Source = "Citizen Science")  
fknms\_add\_source <- fknms\_fix\_names %>% mutate(Source = "FKNMS")  
  
cs\_fknms\_combined <- rbind(cs\_add\_source, fknms\_add\_source)  
  
cs\_fknms\_combined$location <- as.factor(as.character(cs\_fknms\_combined$location))  
cs\_fknms\_combined$Source <- as.factor(as.character(cs\_fknms\_combined$Source))

Now we can start to look at some plots for comparison.We will use side by side boxplots to view each parameter for all of the sites to visualize data source patterns.

effic\_water\_temp <- ggplot(cs\_fknms\_combined, aes(x = location, y = water\_temp, fill = Source)) + geom\_boxplot() + theme(axis.text.x = element\_text(angle = 45, hjust = 1)) + xlab("Sampling Location") + ylab("Water Temperature (C)") + ggtitle("Citizen Science vs. FKNMS Water Temperature") + labs(fill = "Data Source")  
  
effic\_water\_temp

## Warning: Removed 5 rows containing non-finite values (stat\_boxplot).



ggsave(filename = "../../results/Efficacy\_Figures/effic\_water\_temp.png", plot = effic\_water\_temp)

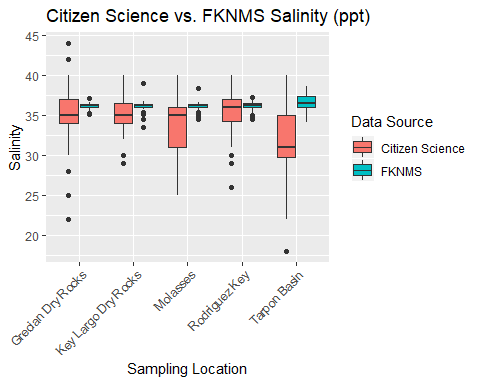
## Saving 5 x 4 in image

## Warning: Removed 5 rows containing non-finite values (stat\_boxplot).

Looking at themperature it looks like there is relatively decent agreement across the boxplots for both mean temperature and interquartile range. There is some small disagreement with the median for FKNMS Rodriguez Key.

effic\_sal <- ggplot(cs\_fknms\_combined, aes(x = location, y = salinity, fill = Source)) + geom\_boxplot() + theme(axis.text.x = element\_text(angle = 45, hjust = 1)) + xlab("Sampling Location") + ylab("Salinity") + ggtitle("Citizen Science vs. FKNMS Salinity (ppt)") + labs(fill = "Data Source")  
  
effic\_sal

## Warning: Removed 9 rows containing non-finite values (stat\_boxplot).



ggsave(filename = "../../results/Efficacy\_Figures/effic\_sal.png", plot = effic\_sal)

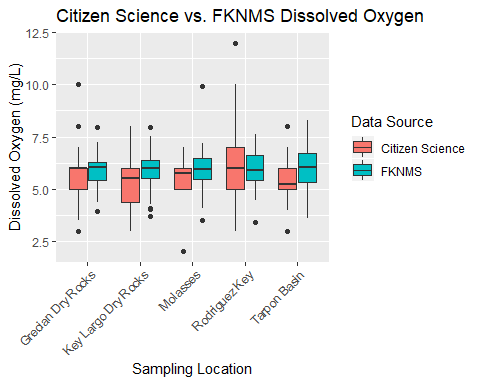
## Saving 5 x 4 in image

## Warning: Removed 9 rows containing non-finite values (stat\_boxplot).

Looking at salinity there are some very interesting patterns in the data. First and formost, it is obvious that the interquartile range for the FKNMS data is much smaller than the citizen science data. This is very likely due to the percision differences of the techniques used to measure water quality between data sources. Additionally, it appears that the citizen science data tends to underestimate the salinity in general as most of the upper limits of the interquartile ranges for citizen science data are similar to the median value of FKNMS data.

effic\_do <- ggplot(cs\_fknms\_combined, aes(x = location, y = dissolved\_oxygen, fill = Source)) + geom\_boxplot() + theme(axis.text.x = element\_text(angle = 45, hjust = 1)) + xlab("Sampling Location") + ylab("Dissolved Oxygen (mg/L)") + ggtitle("Citizen Science vs. FKNMS Dissolved Oxygen") + labs(fill = "Data Source")  
  
effic\_do

## Warning: Removed 4 rows containing non-finite values (stat\_boxplot).



ggsave(filename = "../../results/Efficacy\_Figures/effic\_do.png", plot = effic\_do)

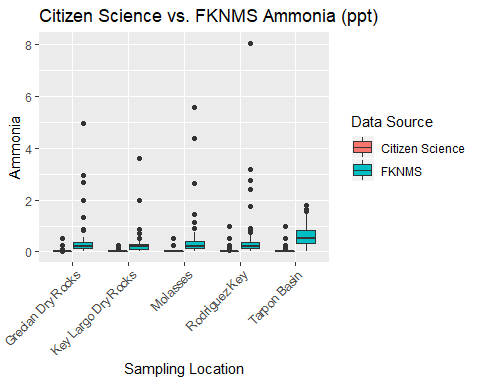
## Saving 5 x 4 in image

## Warning: Removed 4 rows containing non-finite values (stat\_boxplot).

The dissolved oxygen data captures a similar trend across both data sets. Coral reef site locations (Grecian, Key Largo Dry Rocks, and Molasses) appear to be reasonably consistent however, it does appear that citizen science overestimates the DO level a small amount. The Seagrass/Mangrove site (Tarpon Basin) shows an unserestimation of the DO level in citizen science data and the Patch Reef/Hardbottom site (Rodriguez Key) appears relatively consistent. Notably the outliers in the data appear to be substantially higher or lower in citizen science data compared to FKNMS data, this may be due to citizen science experiece-based error.

effic\_amm <- ggplot(cs\_fknms\_combined, aes(x = location, y = ammonia, fill = Source)) + geom\_boxplot() + theme(axis.text.x = element\_text(angle = 45, hjust = 1)) + xlab("Sampling Location") + ylab("Ammonia") + ggtitle("Citizen Science vs. FKNMS Ammonia (ppt)") + labs(fill = "Data Source")  
  
effic\_amm

## Warning: Removed 2 rows containing non-finite values (stat\_boxplot).



ggsave(filename = "../../results/Efficacy\_Figures/effic\_amm.png", plot = effic\_amm)

## Saving 5 x 4 in image

## Warning: Removed 2 rows containing non-finite values (stat\_boxplot).

Ammonia shows an interesting pattern as well. It appeas that the citizen science data may underestimate the total ammonia in the water when compared to FKNMS. Similar to salinity, this is likely due to the measurement specificty of the data collection tools. Additionally, FKNMS outlier data appears to be substantially higher than any citizen science outlier data.

As a final comparison, we will use a Welch’s two-sample t-test to compare the FKNMS findings with the citizen science findings to see if there are any significant differences in the data between the two organizations.

First we must make subsets of the data for each site of interest for comparison.

We will start with Grecian Dry Rocks.

grecian\_only\_cs <- subset(cs\_total, location == "Grecian Dry Rocks")  
grecian\_only\_fknms <- subset(fknms\_fix\_names, location == "Grecian Dry Rocks")  
  
t.test(grecian\_only\_cs$water\_temp, grecian\_only\_fknms$water\_temp)

##   
## Welch Two Sample t-test  
##   
## data: grecian\_only\_cs$water\_temp and grecian\_only\_fknms$water\_temp  
## t = -0.74521, df = 117.77, p-value = 0.4576  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -1.4257933 0.6461148  
## sample estimates:  
## mean of x mean of y   
## 26.09714 26.48698

t.test(grecian\_only\_cs$salinity, grecian\_only\_fknms$salinity)

##   
## Welch Two Sample t-test  
##   
## data: grecian\_only\_cs$salinity and grecian\_only\_fknms$salinity  
## t = -2.2672, df = 70.107, p-value = 0.02646  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -1.8304060 -0.1171964  
## sample estimates:  
## mean of x mean of y   
## 35.16667 36.14047

t.test(grecian\_only\_cs$dissolved\_oxygen, grecian\_only\_fknms$dissolved\_oxygen)

##   
## Welch Two Sample t-test  
##   
## data: grecian\_only\_cs$dissolved\_oxygen and grecian\_only\_fknms$dissolved\_oxygen  
## t = -1.5404, df = 120.16, p-value = 0.1261  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.6355554 0.0793405  
## sample estimates:  
## mean of x mean of y   
## 5.626761 5.904868

t.test(grecian\_only\_cs$ammonia, grecian\_only\_fknms$ammonia)

##   
## Welch Two Sample t-test  
##   
## data: grecian\_only\_cs$ammonia and grecian\_only\_fknms$ammonia  
## t = -3.7501, df = 56.919, p-value = 0.0004158  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.6493195 -0.1972529  
## sample estimates:  
## mean of x mean of y   
## 0.04169014 0.46497635

Next we will compare Key Largo Dry Rocks

kldr\_only\_cs <- subset(cs\_total, location == "Key Largo Dry Rocks")  
kldr\_only\_fknms <- subset(fknms\_fix\_names, location == "Key Largo Dry Rocks")  
  
t.test(kldr\_only\_cs$water\_temp, kldr\_only\_fknms$water\_temp)

##   
## Welch Two Sample t-test  
##   
## data: kldr\_only\_cs$water\_temp and kldr\_only\_fknms$water\_temp  
## t = 0.01107, df = 61.172, p-value = 0.9912  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -1.138727 1.151406  
## sample estimates:  
## mean of x mean of y   
## 26.73548 26.72914

t.test(kldr\_only\_cs$salinity, kldr\_only\_fknms$salinity)

##   
## Welch Two Sample t-test  
##   
## data: kldr\_only\_cs$salinity and kldr\_only\_fknms$salinity  
## t = -1.6371, df = 31.317, p-value = 0.1116  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -1.9924696 0.2176794  
## sample estimates:  
## mean of x mean of y   
## 35.22581 36.11320

t.test(kldr\_only\_cs$dissolved\_oxygen, kldr\_only\_fknms$dissolved\_oxygen)

##   
## Welch Two Sample t-test  
##   
## data: kldr\_only\_cs$dissolved\_oxygen and kldr\_only\_fknms$dissolved\_oxygen  
## t = -2.4695, df = 45.149, p-value = 0.01738  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -1.0711551 -0.1088363  
## sample estimates:  
## mean of x mean of y   
## 5.312500 5.902496

t.test(kldr\_only\_cs$ammonia, kldr\_only\_fknms$ammonia)

##   
## Welch Two Sample t-test  
##   
## data: kldr\_only\_cs$ammonia and kldr\_only\_fknms$ammonia  
## t = -4.2599, df = 68.787, p-value = 6.364e-05  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.3938436 -0.1426080  
## sample estimates:  
## mean of x mean of y   
## 0.02727273 0.29549853

Next we compare Molasses.

mo\_only\_cs <- subset(cs\_total, location == "Molasses")  
mo\_only\_fknms <- subset(fknms\_fix\_names, location == "Molasses")  
  
t.test(mo\_only\_cs$water\_temp, mo\_only\_fknms$water\_temp)

##   
## Welch Two Sample t-test  
##   
## data: mo\_only\_cs$water\_temp and mo\_only\_fknms$water\_temp  
## t = 0.3961, df = 20.38, p-value = 0.6962  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -1.311654 1.927453  
## sample estimates:  
## mean of x mean of y   
## 26.96429 26.65639

t.test(mo\_only\_cs$salinity, mo\_only\_fknms$salinity)

##   
## Welch Two Sample t-test  
##   
## data: mo\_only\_cs$salinity and mo\_only\_fknms$salinity  
## t = -1.4844, df = 13.072, p-value = 0.1614  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -4.450989 0.824334  
## sample estimates:  
## mean of x mean of y   
## 34.28571 36.09904

t.test(mo\_only\_cs$dissolved\_oxygen, mo\_only\_fknms$dissolved\_oxygen)

##   
## Welch Two Sample t-test  
##   
## data: mo\_only\_cs$dissolved\_oxygen and mo\_only\_fknms$dissolved\_oxygen  
## t = -1.4968, df = 16.874, p-value = 0.1529  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -1.2098830 0.2059671  
## sample estimates:  
## mean of x mean of y   
## 5.428571 5.930529

t.test(mo\_only\_cs$ammonia, mo\_only\_fknms$ammonia)

##   
## Welch Two Sample t-test  
##   
## data: mo\_only\_cs$ammonia and mo\_only\_fknms$ammonia  
## t = -3.5322, df = 74.897, p-value = 0.0007094  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.6561530 -0.1829236  
## sample estimates:  
## mean of x mean of y   
## 0.05357143 0.47310972

Next we compare Rodriguez Key.

roddy\_only\_cs <- subset(cs\_total, location == "Rodriguez Key")  
roddy\_only\_fknms <- subset(fknms\_fix\_names, location == "Rodriguez Key")  
  
t.test(roddy\_only\_cs$water\_temp, roddy\_only\_fknms$water\_temp)

##   
## Welch Two Sample t-test  
##   
## data: roddy\_only\_cs$water\_temp and roddy\_only\_fknms$water\_temp  
## t = 2.8276, df = 119.99, p-value = 0.005498  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 0.5803274 3.2914327  
## sample estimates:  
## mean of x mean of y   
## 27.76615 25.83027

t.test(roddy\_only\_cs$salinity, roddy\_only\_fknms$salinity)

##   
## Welch Two Sample t-test  
##   
## data: roddy\_only\_cs$salinity and roddy\_only\_fknms$salinity  
## t = -1.0301, df = 69.395, p-value = 0.3065  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -1.1103294 0.3540844  
## sample estimates:  
## mean of x mean of y   
## 35.72727 36.10540

t.test(roddy\_only\_cs$dissolved\_oxygen, roddy\_only\_fknms$dissolved\_oxygen)

##   
## Welch Two Sample t-test  
##   
## data: roddy\_only\_cs$dissolved\_oxygen and roddy\_only\_fknms$dissolved\_oxygen  
## t = 0.7519, df = 97.895, p-value = 0.4539  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.2788176 0.6189837  
## sample estimates:  
## mean of x mean of y   
## 6.130769 5.960686

t.test(roddy\_only\_cs$ammonia, roddy\_only\_fknms$ammonia)

##   
## Welch Two Sample t-test  
##   
## data: roddy\_only\_cs$ammonia and roddy\_only\_fknms$ammonia  
## t = -3.0701, df = 65.365, p-value = 0.003115  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.7371541 -0.1561267  
## sample estimates:  
## mean of x mean of y   
## 0.06507692 0.51171734

And finally we compare Tarpon Basin.

tarp\_only\_cs <- subset(cs\_total, location == "Tarpon Basin")  
tarp\_only\_fknms <- subset(fknms\_fix\_names, location == "Tarpon Basin")  
  
t.test(tarp\_only\_cs$water\_temp, tarp\_only\_fknms$water\_temp)

##   
## Welch Two Sample t-test  
##   
## data: tarp\_only\_cs$water\_temp and tarp\_only\_fknms$water\_temp  
## t = -0.59978, df = 116.16, p-value = 0.5498  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -1.968509 1.053393  
## sample estimates:  
## mean of x mean of y   
## 26.89833 27.35589

t.test(tarp\_only\_cs$salinity, tarp\_only\_fknms$salinity)

##   
## Welch Two Sample t-test  
##   
## data: tarp\_only\_cs$salinity and tarp\_only\_fknms$salinity  
## t = -8.8835, df = 60.896, p-value = 1.369e-12  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -6.604946 -4.177748  
## sample estimates:  
## mean of x mean of y   
## 31.16964 36.56099

t.test(tarp\_only\_cs$dissolved\_oxygen, tarp\_only\_fknms$dissolved\_oxygen)

##   
## Welch Two Sample t-test  
##   
## data: tarp\_only\_cs$dissolved\_oxygen and tarp\_only\_fknms$dissolved\_oxygen  
## t = -2.2926, df = 111.5, p-value = 0.02375  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.89358157 -0.06504758  
## sample estimates:  
## mean of x mean of y   
## 5.486949 5.966264

t.test(tarp\_only\_cs$ammonia, tarp\_only\_fknms$ammonia)

##   
## Welch Two Sample t-test  
##   
## data: tarp\_only\_cs$ammonia and tarp\_only\_fknms$ammonia  
## t = -9.3754, df = 77.57, p-value = 2.125e-14  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.6652810 -0.4322103  
## sample estimates:  
## mean of x mean of y   
## 0.06583333 0.61457901

Based on the overall comparison of both datasets, the efficacy of citizen science data is situationally dependant. Viewing the water quality comparison boxplots, the citizen science data captured the major trends in water quality of all of the comparable sites. However, it is evident by the visualizations that the methods employed by the FKNMS team are more accurate and percise than those used by the citizen scientists. This result is not suprising due to the fact that citizen science data collection in itself is intended largely for the purpose of monitoring and therefore tools are typically less sensetive and cheaper.

Viewing the results of the t-test we see that ammonia was significantly different between the two sites in all instances. This suggests that there may be a need to improve the tools used to measure ammonia levels by the citizen scientists, however this may be a limiting resource dependant on funding. Salinity was significant in two of the three locations, this is likely due to the differences in the equipment senestivity of the two agencies however this significance was not consistent across all measures. This result suggests that while sainily measures from FKNMS are more accurate and percise the citizen science measures still capture the general trends of the salinity with some biases introduced by equipment specificity. Of the remaining paramaters, significance was present in at least one of the samples for each variable, however there is no distinct pattern of significance for dissolved oxygen and temperature seen for all sites. This result suggests that this significance is likely due to an confounding factor (season, specific location at the site, time of day, etc.) rather than the usage of citicen science vs. FKNMS data.

Regarding the efficacy of the data, we conclude that the utlity of citizen science data is dependant on the overall goal of the research. If a research study aims to capture large-scale trends and/or general patterns of water quality, citizen science data collection is a viable option. However, if a study is interested in specific details and/or very accurate measures of water quality parameters this methods is not ideal for data collection.