WQProcessing\_cleaning

William Norfolk

10/7/2019

Load needed packages. make sure they are installed.

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 0.3.2  
## v tidyr 0.8.3 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)

Load the data. This path is relative to the processinf code folder in the project directory.

wqrawdata <- readxl::read\_excel("../../data/raw\_data/RAW\_WQ Data.xlsx")

## New names:  
## \* `` -> ...13

Take a look at the data.

glimpse(wqrawdata)

## Observations: 522  
## Variables: 13  
## $ Timestamp <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...  
## $ Date <chr> "42377", "42408", "42408", "42408", "42409"...  
## $ Time <chr> "0.59375", "0.63541666666666663", "0.659722...  
## $ Location <chr> "Largo Sound", "Grecian Dry Rocks", "Grecia...  
## $ `Instructor Name` <chr> "Katy, Sarah, Driver", "Chelsea", "Katy, To...  
## $ `Group Name` <chr> "NA", "McLean HS", "McLean HS", "McLean HS"...  
## $ pH <chr> "8", "8.4", "8.1999999999999993", "8.4", "8...  
## $ Ammonia <chr> "0", "0", "0", "0", "0.25", "0", "0", "0", ...  
## $ `Dissolved Oxygen` <chr> "5", "4", "4", "6", "8", "4", "5", "6", "6"...  
## $ `Water Temp (C)` <chr> "NA", "23.5", "21", "36", "18", "18", "18",...  
## $ `Salinity (ppt)` <chr> "36", "40", "44", "35", "33", "30", "33", "...  
## $ Equipment <chr> "kit", "kit", "kit", "kit", "kit", "kit", "...  
## $ ...13 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...

At a glance there is clearly some cleaning to be done. First the values for the variables “Date” and “Time” did not read into r properly. This is likely due to the autoformatting of dates by Excel. These data will not read into the program properly and will cause problems downstream so it is best to solve this issue by reformatting the data entries in Excel by adding new columns to replace the date and time variables. Date will now be represented as the new variable “Numeric Date” which has reformatted the date to a mm.dd.yy style. Time will be represented as the new variable “Military Time” which has reformatted the time to standard numeric military time. Both original variables will be kept in the raw data set, but will be removed by further r processing.

Reload the data with the reformatted Excel sheet.

wqrawdata <- readxl::read\_excel("../../data/raw\_data/RAW\_WQ DATA\_TimeModified.xlsx")

## New names:  
## \* `` -> ...15

Take a look at the data once more.

glimpse(wqrawdata)

## Observations: 522  
## Variables: 15  
## $ Timestamp <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...  
## $ `Numeric Date` <chr> "01.08.16", "02.08.16", "02.08.16", "02.08....  
## $ Date <chr> "42377", "42408", "42408", "42408", "42409"...  
## $ Time <chr> "0.59375", "0.63541666666666663", "0.659722...  
## $ `Military Time` <chr> "1415", "1515", "1550", "1555", "1001", "10...  
## $ Location <chr> "Largo Sound", "Grecian Dry Rocks", "Grecia...  
## $ `Instructor Name` <chr> "Katy, Sarah, Driver", "Chelsea", "Katy, To...  
## $ `Group Name` <chr> "NA", "McLean HS", "McLean HS", "McLean HS"...  
## $ pH <chr> "8", "8.4", "8.1999999999999993", "8.4", "8...  
## $ Ammonia <chr> "0", "0", "0", "0", "0.25", "0", "0", "0", ...  
## $ `Dissolved Oxygen` <chr> "5", "4", "4", "6", "8", "4", "5", "6", "6"...  
## $ `Water Temp (C)` <chr> "NA", "23.5", "21", "36", "18", "18", "18",...  
## $ `Salinity (ppt)` <chr> "36", "40", "44", "35", "33", "30", "33", "...  
## $ Equipment <chr> "kit", "kit", "kit", "kit", "kit", "kit", "...  
## $ ...15 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...

We should have 11 variables in our data set but it seems that 15 have been loaded. The variables “Timestamp” and “…15” are errors and should be dropped. Also the old “Date” and “Time” variables should be removed.

wqrawdata\_11 <- select(wqrawdata, -c(Timestamp, ...15, Date, Time))  
  
glimpse(wqrawdata\_11)

## Observations: 522  
## Variables: 11  
## $ `Numeric Date` <chr> "01.08.16", "02.08.16", "02.08.16", "02.08....  
## $ `Military Time` <chr> "1415", "1515", "1550", "1555", "1001", "10...  
## $ Location <chr> "Largo Sound", "Grecian Dry Rocks", "Grecia...  
## $ `Instructor Name` <chr> "Katy, Sarah, Driver", "Chelsea", "Katy, To...  
## $ `Group Name` <chr> "NA", "McLean HS", "McLean HS", "McLean HS"...  
## $ pH <chr> "8", "8.4", "8.1999999999999993", "8.4", "8...  
## $ Ammonia <chr> "0", "0", "0", "0", "0.25", "0", "0", "0", ...  
## $ `Dissolved Oxygen` <chr> "5", "4", "4", "6", "8", "4", "5", "6", "6"...  
## $ `Water Temp (C)` <chr> "NA", "23.5", "21", "36", "18", "18", "18",...  
## $ `Salinity (ppt)` <chr> "36", "40", "44", "35", "33", "30", "33", "...  
## $ Equipment <chr> "kit", "kit", "kit", "kit", "kit", "kit", "...

The resulting data set is a tibble with 522 observations and 11 variables. Next the correct classes must be assigned to each variable so the data can be properly manipulated. All variables are currently listed as “characters” we must change Military Time, pH, Dissolved Oxygen, Ammonia, Water Temp, and Salinity to numeric classes.

A few of our variables have been introduced into R with a few extra bits (some ` and units for salinity and temperature) and spaces between words rather than underscores. These will likely cause problems and/or frequent typos downstream so we will recode the variables to a more R-friendly format.

wqraw\_reformat <- wqrawdata\_11 %>% rename(numeric\_date = `Numeric Date`,  
 military\_time = `Military Time`,  
 location = Location,  
 instructor\_name = `Instructor Name`,  
 group\_name = `Group Name`,  
 ph = pH,  
 ammonia = Ammonia,  
 dissolved\_oxygen = `Dissolved Oxygen`,  
 water\_temp = `Water Temp (C)`,  
 salinity = `Salinity (ppt)`,  
 equipment = Equipment)  
  
glimpse(wqraw\_reformat)

## Observations: 522  
## Variables: 11  
## $ numeric\_date <chr> "01.08.16", "02.08.16", "02.08.16", "02.08.16...  
## $ military\_time <chr> "1415", "1515", "1550", "1555", "1001", "1015...  
## $ location <chr> "Largo Sound", "Grecian Dry Rocks", "Grecian ...  
## $ instructor\_name <chr> "Katy, Sarah, Driver", "Chelsea", "Katy, Tomm...  
## $ group\_name <chr> "NA", "McLean HS", "McLean HS", "McLean HS", ...  
## $ ph <chr> "8", "8.4", "8.1999999999999993", "8.4", "8",...  
## $ ammonia <chr> "0", "0", "0", "0", "0.25", "0", "0", "0", "0...  
## $ dissolved\_oxygen <chr> "5", "4", "4", "6", "8", "4", "5", "6", "6", ...  
## $ water\_temp <chr> "NA", "23.5", "21", "36", "18", "18", "18", "...  
## $ salinity <chr> "36", "40", "44", "35", "33", "30", "33", "35...  
## $ equipment <chr> "kit", "kit", "kit", "kit", "kit", "kit", "ki...

Next we will need to convert our variables from their loaded format to a working formay. Variables Military Time, pH, Ammoina, Dissolved Oxygen, Water Temp (C), and Salinity (ppt) should be numeric so they can be used for exploratory analysis.

wqraw\_reformat$military\_time <- as.numeric(as.character(wqraw\_reformat$military\_time))

## Warning: NAs introduced by coercion

wqraw\_reformat$ph <- as.numeric(as.character(wqraw\_reformat$ph))

## Warning: NAs introduced by coercion

wqraw\_reformat$ammonia <- as.numeric(as.character(wqraw\_reformat$ammonia))

## Warning: NAs introduced by coercion

wqraw\_reformat$dissolved\_oxygen <- as.numeric(as.character(wqraw\_reformat$dissolved\_oxygen))

## Warning: NAs introduced by coercion

wqraw\_reformat$water\_temp <- as.numeric(as.character(wqraw\_reformat$water\_temp))

## Warning: NAs introduced by coercion

wqraw\_reformat$salinity <- as.numeric(as.character(wqraw\_reformat$salinity))

## Warning: NAs introduced by coercion

glimpse(wqraw\_reformat)

## Observations: 522  
## Variables: 11  
## $ numeric\_date <chr> "01.08.16", "02.08.16", "02.08.16", "02.08.16...  
## $ military\_time <dbl> 1415, 1515, 1550, 1555, 1001, 1015, 1022, 103...  
## $ location <chr> "Largo Sound", "Grecian Dry Rocks", "Grecian ...  
## $ instructor\_name <chr> "Katy, Sarah, Driver", "Chelsea", "Katy, Tomm...  
## $ group\_name <chr> "NA", "McLean HS", "McLean HS", "McLean HS", ...  
## $ 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...

Next, we will want to compare sites that were located on the Bayside (Gulf of Mexico side) of the island to sites on the Oceanside of the island (Atlantic Ocean). There is site specific data for each location, however bay vs. ocean is not specified so we will need to add a new variable based on the specific site data.

First take a look at all fo the unique responses for location to see what is present and if any cleaning is needed.

unique(wqraw\_reformat$location)

## [1] "Largo Sound" "Grecian Dry Rocks"   
## [3] "NA" "Tarpon Basin"   
## [5] "Nest Key" "Dana's Cove"   
## [7] "Pelican Key" "Horseshoe"   
## [9] "Higdon's" "Rodriguez Key"   
## [11] "Key Largo Dry Rocks" "Elbow Nursery"   
## [13] "Molasses" "Sexton Cove"   
## [15] "The Barge" "Scott's Channel"   
## [17] "Radabob" "Bayside"   
## [19] "Minnow Caves" "Banana"   
## [21] "Blackwater Sound" "Cannon Patch"   
## [23] "Sexton Cov" "Whitmore bite"   
## [25] "Bush Point" "Captain's Point"   
## [27] "North Dry Rocks/Minnow Caves" "Ocean Reef Club"   
## [29] "Whitmore Bite" "Benny Poo"   
## [31] "Grotto" "North North"   
## [33] "El Radabob" "Higdons"   
## [35] "Swash Key" "Dry Rocks"   
## [37] "Rodriguez" "The Grotto"   
## [39] "Plane Wreck" "French Reef"   
## [41] "Sexton" "Whitmore Bight"   
## [43] "Canon Patch" "Boat Ramp"   
## [45] "Whitmore" "Deep Blue (off molasses)"   
## [47] "Swash Keys" "Buttonwood"   
## [49] "Marker 37" "North Dry Rocks"   
## [51] "Danas Cove" "whitmore"   
## [53] "Hardbottom" "Lagoon"   
## [55] "The Elbow" "Deep Blue (Elbow)"   
## [57] "Lakewatch 2" "Tarpon Deep"   
## [59] "Rodriguez Key (MG/SG Side)" "Tarpon Basin (Tarpon Island)"  
## [61] "South Sound Creek" "Hidden Lake"   
## [63] "Higdens" "key Largo Dry Rocks"   
## [65] "North North Dry Rocks" "Mosquito"   
## [67] "Molasses Reef" "Rodriguez Key (North)"   
## [69] "Elbow" "Tarpon Island"

There is definitely some disagreement with the correct site names and spelling so we will have to correct these before moving forward. A number of the locations are showing a simple typo in entry, or the instructor entered a different variant of a synonym for the site so we will correct with the code below.

recodelocation <- wqraw\_reformat %>% mutate(location = recode(location,   
 "Largo Sound" = "Boat Ramp",  
 "Higdens" = "Higdon's",  
 "Higdons" = "Higdon's",  
 "Dry Rocks" = "Key Largo Dry Rocks",  
 "key Largo Dry Rocks" = "Key Largo Dry Rocks",  
 "Molasses Reef" = "Molasses",  
 "Radabob" = "El Radabob",  
 "North Dry Rocks" = "Minnow Caves",  
 "North Dry Rocks/Minnow Caves" = "Minnow Caves",  
 "Sexton Cov" = "Sexton Cove",  
 "Sexton" = "Sexton Cove",  
 "Whitmore Bite" = "Whitmore Bight",  
 "Whitmore" = "Whitmore Bight",  
 "Hardbottom" = "Whitmore Bight",  
 "Whitmore bite" = "Whitmore Bight",  
 "whitmore" = "Whitmore Bight",  
 "Grotto" = "The Grotto",  
 "Swash Key" = "Swash Keys",  
 "Plane Wreck" = "Rodriguez Key",  
 "Rodriguez" = "Rodriguez Key",  
 "Rodriguez Key (MG/SG Side)" = "Rodriguez Key",  
 "Rodriguez Key (North)" = "Rodriguez Key",  
 "Canon Patch" = "Cannon Patch",  
 "Danas Cove" = "Dana's Cove",  
 "Elbow" = "The Elbow",  
 "Elbow Nursery" = "The Elbow",  
 "Lakewatch 2" = "Tarpon Basin",  
 "North North" = "North North Dry Rocks",  
 "Deep Blue (off molasses)" = "Deep Blue",  
 "Deep Blue (Elbow)" = "Deep Blue",  
 "Tarpon Basin (Tarpon Island)" = "Tarpon Island"))  
  
unique(recodelocation$location)

## [1] "Boat Ramp" "Grecian Dry Rocks"   
## [3] "NA" "Tarpon Basin"   
## [5] "Nest Key" "Dana's Cove"   
## [7] "Pelican Key" "Horseshoe"   
## [9] "Higdon's" "Rodriguez Key"   
## [11] "Key Largo Dry Rocks" "The Elbow"   
## [13] "Molasses" "Sexton Cove"   
## [15] "The Barge" "Scott's Channel"   
## [17] "El Radabob" "Bayside"   
## [19] "Minnow Caves" "Banana"   
## [21] "Blackwater Sound" "Cannon Patch"   
## [23] "Whitmore Bight" "Bush Point"   
## [25] "Captain's Point" "Ocean Reef Club"   
## [27] "Benny Poo" "The Grotto"   
## [29] "North North Dry Rocks" "Swash Keys"   
## [31] "French Reef" "Deep Blue"   
## [33] "Buttonwood" "Marker 37"   
## [35] "Lagoon" "Tarpon Deep"   
## [37] "Tarpon Island" "South Sound Creek"   
## [39] "Hidden Lake" "Mosquito"

Now that there is no overlap with the site names we can create a new variable that determines if an observation is in the bayside or oceanside.

add\_bay\_or\_ocean <- recodelocation %>% mutate(island\_side =   
 ifelse(location == "Rodriguez Key", "ocean",  
 ifelse(location == "Molasses", "ocean",  
 ifelse(location == "Minnow Caves", "ocean",  
 ifelse(location == "Cannon Patch", "ocean",  
 ifelse(location == "French Reef", "ocean",  
 ifelse(location == "Marker 37", "ocean",  
 ifelse(location == "Mosquito", "ocean",  
 ifelse(location == "Grecian Dry Rocks", "ocean",  
 ifelse(location == "Horseshoe", "ocean",  
 ifelse(location == "Key Largo Dry Rocks", "ocean",  
 ifelse(location == "El Radabob", "ocean",  
 ifelse(location == "Banana", "ocean",  
 ifelse(location == "Whitmore Bight", "ocean",  
 ifelse(location == "Ocean Reef Club", "ocean",  
 ifelse(location == "North North Dry Rocks", "ocean",  
 ifelse(location == "Deep Blue", "ocean",  
 ifelse(location == "Lagoon", "ocean",  
 ifelse(location == "South Sound Creek", "ocean",   
 ifelse(location == "Higdon's", "ocean",  
 ifelse(location == "The Elbow", "ocean",   
 ifelse(location == "The Barge", "ocean",   
 ifelse(location == "Benny Poo", "ocean",  
 ifelse(location == "Boat Ramp", "ocean",  
 ifelse(location == "Tarpon Basin", "bay",  
 ifelse(location == "Pelican Key", "bay",  
 ifelse(location == "Scott's Channel", "bay",  
 ifelse(location == "Captain's Point", "bay",   
 ifelse(location == "The Grotto", "bay",   
 ifelse(location == "Tarpon Island", "bay",   
 ifelse(location == "Nest Key", "bay",  
 ifelse(location == "Sexton Cove", "bay",   
 ifelse(location == "Dana's Cove", "bay",   
 ifelse(location == "Bayside", "bay",   
 ifelse(location == "Blackwater Sound", "bay",   
 ifelse(location == "Bush Point", "bay",   
 ifelse(location == "Swash Keys", "bay",   
 ifelse(location == "Buttonwood", "bay",   
 ifelse(location == "Tarpon Deep", "bay",   
 ifelse(location == "Hidden Lake", "bay",   
 "no"))))))))))))))))))))))))))))))))))))))))

It will also be useful to know which type of site we are looking at with out data. Of the locations visited there are three different categories of sites: Seagrass/Mangroves, Coral Reefs, and Patch Reef/Hardbottom. Each of these types of ecosystems have characteristic water quality differences which may be useful for comparison for the full analysis we we will code them in as well using similar code to above.

add\_site\_type <- add\_bay\_or\_ocean %>% mutate(site\_type =   
 ifelse(location == "Rodriguez Key", "Patch Reef/Hardbottom",  
 ifelse(location == "Molasses", "Coral Reef",  
 ifelse(location == "Minnow Caves", "Coral Reef",  
 ifelse(location == "Cannon Patch", "Patch Reef/Hardbottom",  
 ifelse(location == "French Reef", "Coral Reef",  
 ifelse(location == "Marker 37", "Patch Reef/Hardbottom",  
 ifelse(location == "Mosquito", "Patch Reef/Hardbottom",  
 ifelse(location == "Grecian Dry Rocks", "Coral Reef",  
 ifelse(location == "Horseshoe", "Coral Reef",  
 ifelse(location == "Key Largo Dry Rocks", "Coral Reef",  
 ifelse(location == "El Radabob", "Seagrass/Mangrove",  
 ifelse(location == "Banana", "Coral Reef",  
 ifelse(location == "Whitmore Bight", "Patch Reef/Hardbottom",  
 ifelse(location == "Ocean Reef Club", "Coral Reef",  
 ifelse(location == "North North Dry Rocks", "Coral Reef",  
 ifelse(location == "Deep Blue", "Coral Reef",  
 ifelse(location == "Lagoon", "Seagrass/Mangrove",  
 ifelse(location == "South Sound Creek", "Seagrass/Mangrove",   
 ifelse(location == "Higdon's", "Patch Reef/Hardbottom",  
 ifelse(location == "The Elbow", "Coral Reef",   
 ifelse(location == "The Barge", "Patch Reef/Hardbottom",   
 ifelse(location == "Benny Poo", "Seagrass/Mangrove",  
 ifelse(location == "Boat Ramp", "Seagrass/Mangrove",  
 ifelse(location == "Tarpon Basin", "Seagrass/Mangrove",  
 ifelse(location == "Pelican Key", "Seagrass/Mangrove",  
 ifelse(location == "Scott's Channel", "Seagrass/Mangrove",  
 ifelse(location == "Captain's Point", "Seagrass/Mangrove",   
 ifelse(location == "The Grotto", "Seagrass/Mangrove",   
 ifelse(location == "Tarpon Island", "Seagrass/Mangrove",   
 ifelse(location == "Nest Key", "Seagrass/Mangrove",  
 ifelse(location == "Sexton Cove", "Seagrass/Mangrove",   
 ifelse(location == "Dana's Cove", "Seagrass/Mangrove",   
 ifelse(location == "Bayside", "Seagrass/Mangrove",   
 ifelse(location == "Blackwater Sound", "Seagrass/Mangrove",   
 ifelse(location == "Bush Point", "Seagrass/Mangrove",   
 ifelse(location == "Swash Keys", "Seagrass/Mangrove",   
 ifelse(location == "Buttonwood", "Seagrass/Mangrove",   
 ifelse(location == "Tarpon Deep", "Seagrass/Mangrove",   
 ifelse(location == "Hidden Lake", "Seagrass/Mangrove",   
 "no"))))))))))))))))))))))))))))))))))))))))

Next lets be sure to change any “no” entries to NA so we can remove from plots with ease.

add\_site\_type[add\_site\_type == "no"] <- NA

Now that we have all of our proper variables to compare the data,let’s check the numeric variables to see if there are any observations that need addressing.

pH data looks good!

unique(add\_site\_type$ph)

## [1] 8.0 8.4 8.2 8.9 8.3 8.1 NA 7.8 7.5 7.9 8.5 8.6 7.4

Some of these ammoina measures seem a bit high. Typically ammonia does not get over 1.0 unless there is severe contamination (none of which was noted during this time). 34.0, 5.0, and 4.0 are too high to be possible for our standard conditions so, we will impute this value from the mean of the other observations.

unique(add\_site\_type$ammonia)

## [1] 0.000 0.250 NA 0.300 0.100 5.000 0.150 0.010 0.125 0.500  
## [11] 1.000 0.130 0.050 4.000 34.000 0.200

mean(add\_site\_type$ammonia, na.rm = TRUE)

## [1] 0.1298633

add\_site\_type$ammonia[add\_site\_type$ammonia == "34"] <- "0.046"  
add\_site\_type$ammonia[add\_site\_type$ammonia == "5"] <- "0.046"  
add\_site\_type$ammonia[add\_site\_type$ammonia == "4"] <- "0.046"  
  
class(add\_site\_type$ammonia)

## [1] "character"

Looks like that swap changes the variable back to a character class so convert back to numeric.

add\_site\_type$ammonia <- as.numeric(as.character(add\_site\_type$ammonia))  
  
  
class(add\_site\_type$ammonia)

## [1] "numeric"

Similarly to ammonia, dissolved oxygen also has an impossible outlier of 29 so lets impute the mean as well.

unique(add\_site\_type$dissolved\_oxygen)

## [1] 5.00 4.00 6.00 8.00 9.00 10.00 7.00 4.50 12.00 5.50 NA  
## [12] 3.00 5.10 3.50 6.50 6.15 7.50 5.23 29.00 2.00

mean(add\_site\_type$dissolved\_oxygen, na.rm = TRUE)

## [1] 5.63027

add\_site\_type$dissolved\_oxygen[add\_site\_type$dissolved\_oxygen == "29"] <- "5.58"  
  
class(add\_site\_type$dissolved\_oxygen)

## [1] "character"

add\_site\_type$dissolved\_oxygen <- as.numeric(as.character(add\_site\_type$dissolved\_oxygen))  
  
class(add\_site\_type$dissolved\_oxygen)

## [1] "numeric"

Water temperature looks good!

unique(add\_site\_type$water\_temp)

## [1] NA 23.5 21.0 36.0 18.0 18.3 23.0 18.5 20.1 19.3 22.0 21.9 22.2 22.1  
## [15] 24.0 23.4 22.8 24.3 25.2 25.0 24.9 24.4 24.5 24.2 28.5 25.9 25.3 28.2  
## [29] 27.3 28.3 28.9 26.5 26.6 27.0 24.6 32.1 28.0 29.1 30.2 27.1 29.8 32.2  
## [43] 29.5 31.3 30.6 30.8 33.2 32.3 32.7 30.3 29.6 31.0 29.0 31.2 32.9 30.7  
## [57] 30.5 36.1 32.8 29.4 34.7 32.0 33.8 33.0 30.1 30.0 33.1 31.8 32.4 30.9  
## [71] 31.6 31.1 27.4 26.0 28.8 20.0 23.6 22.5 31.5 34.0 19.0 17.5 27.5 32.5  
## [85] 12.0 20.5

Salinity looks interesting. All site should represent salt or brackish water so we should not see any values below ~10. Upon closer inspection of the data, there are three very low salinity measures all which come from ocean side sites. These are clear errors ad will be imputed from the mean.

unique(add\_site\_type$salinity)

## [1] 36.0 40.0 44.0 35.0 33.0 30.0 25.0 37.0 39.0 31.0 26.0 20.0 22.0 32.0  
## [15] 28.0 24.0 34.0 NA 50.0 0.0 29.0 38.0 18.0 27.0 34.8 48.0 34.5 33.5  
## [29] 2.9 31.5 42.0 30.5 6.0 21.0 35.5 25.5 43.0

mean(add\_site\_type$salinity, na.rm = TRUE)

## [1] 33.94407

add\_site\_type$salinity[add\_site\_type$salinity == "6"] <- "34.1"  
add\_site\_type$salinity[add\_site\_type$salinity == "2.9"] <- "34.1"  
add\_site\_type$salinity[add\_site\_type$salinity == "0"] <- "34.1"  
  
  
class(add\_site\_type$salinity)

## [1] "character"

add\_site\_type$salinity <- as.numeric(as.character(add\_site\_type$salinity))  
  
class(add\_site\_type$salinity)

## [1] "numeric"

Next let’s take a look at the instructor\_name variable. This is not of primary interest to our data analysis of water quality but could be useful for quality control. There is some recoding to do here to account for some name overlaps between upper and lower case, however the majorty of the input looks good.

unique(add\_site\_type$instructor\_name)

## [1] "Katy, Sarah, Driver" "Chelsea"   
## [3] "Katy, Tommy, Lauren" "Breege"   
## [5] "Shaun" "Jessica"   
## [7] "NA" "Lauren"   
## [9] "Driver" "Courtney"   
## [11] "Jessica, Chelsea" "Darien"   
## [13] "Brittany" "Katy"   
## [15] "Sarah" "Tanner"   
## [17] "Breege/Brittany" "Chelsea/Darien"   
## [19] "Jessica/Courtney" "Laure"   
## [21] "Breege/Sarah" "Jessica/Tanner"   
## [23] "Lauren/Breege/Courtney" "Brittany/Tanner"   
## [25] "Billy & Tanner" "Alex"   
## [27] "Billy" "Jessica & Billy"   
## [29] "Jessica & Tanner" "Brittany & Darien"   
## [31] "Rachel" "Anna"   
## [33] "Matt" "Maria"   
## [35] "Pat & Rachel" "Emily"   
## [37] "Maddie" "Pat"   
## [39] "Rachel, Billy, Emily" "Ellery"   
## [41] "Billy, Britt" "Tanner, Maddie, Rachel"  
## [43] "Maddie, Billy" "maddie"   
## [45] "Emma" "maria"   
## [47] "Maria & Maddie" "Maddie & Emma"   
## [49] "Megan" "William"   
## [51] "n/a" "Bill & Matt"   
## [53] "Emma & Nicole" "Megan & Rachel"   
## [55] "Nicole & Rachel" "Cara"   
## [57] "Maddie & Megan" "Billy & Megan"   
## [59] "Anna & Maria" "Nicole"   
## [61] "Bill" "Maddie & Nicole"   
## [63] "Anna & William" "Anna & Megan"   
## [65] "Pat & Nicole" "Megan & Bill"   
## [67] "Pat & Cara" "Rachel & Megan"   
## [69] "Sarah and Cara" "Maria and Pat"   
## [71] "Emma, Megan and Cara" "Rachel and Nicole"   
## [73] "Megan and Emma"

instructor\_recode <- add\_site\_type %>% mutate(instructor\_name = recode(instructor\_name, "maria" = "Maria", "Laure" = "Lauren", "maddie" = "Maddie", "n/a" = "NA"))

The group\_name variable could be relatively interesting as well for some quality control.

unique(instructor\_recode$group\_name)

## [1] "NA" "McLean HS"   
## [3] "Ocean Studies CS" "Three Oaks MS"   
## [5] "South Lakes HS" "Hamilton West"   
## [7] "Hamilton West HS" "BELL"   
## [9] "Episcopal HS" "Lake Braddock"   
## [11] "St. Paul" "James Wood HS"   
## [13] "New Hanover" "New Hanover HS"   
## [15] "Escambia HS" "Mt Vernon"   
## [17] "Barron/Naples" "Naples"   
## [19] "St. Catherine's" "SCM"   
## [21] "Deland HS" "Georgia Military Academy"   
## [23] "GMC HS" "The Burlington School"   
## [25] "TBS" "MPCS"   
## [27] "Woodland" "WPMS"   
## [29] "FIS" "Holy Child"   
## [31] "Two Rivers" "Fairfield Middle"   
## [33] "Fairfield" "RCLS"   
## [35] "Lutheran Schools MW" "The Lab School"   
## [37] "ORC" "Classen"   
## [39] "Classen SAS" "Dive In"   
## [41] "Dive Ia" "TWS Advanced"   
## [43] "TWS" "Heritage Christian"   
## [45] "F.I.N." "UWF"   
## [47] "Georgia Military" "TVM"   
## [49] "Treasure Village Montessori" "Plantation Key"   
## [51] "Plantation Key School" "Seacrest"   
## [53] "Stellar Secondary" "Carmel Christian"   
## [55] "Wheaton Montessori" "Lyme School"   
## [57] "Monarch HS" "One Day Academy"   
## [59] "Vierra HS" "KLS"   
## [61] "Naples / Baron" "NHHS"   
## [63] "CNCC" "Paran"   
## [65] "Mount Paran" "Sackets"   
## [67] "Boynton" "Ocean Reef"   
## [69] "Maret School" "MCSMS"   
## [71] "UW Manitowoc" "Fond du lac"   
## [73] "VW" "MarineLab Phyto"   
## [75] "GMC" "Three Oaks"   
## [77] "Viera" "Stanwich"   
## [79] "PIUS" "Franklin Road Academy"   
## [81] "Rustic Pathways" "Monarch High School"   
## [83] "MAST" "MarineLab"   
## [85] "McClean" "St. Paul Lutheran"   
## [87] "Unity" "Woodland Park"   
## [89] "MRDF" "Ocean First"   
## [91] "NRS" "Alabama school for the deaf"   
## [93] "Lutheran" "Midwest"   
## [95] "Marinelab" "LSMW"   
## [97] "Bell" "Manitowoc"   
## [99] "University of Wisconsin" "Hamilton High"   
## [101] "Treasure Village" "Key Largo School"   
## [103] "Lutz" "Lutz Prep"   
## [105] "Steinert" "South Broward"   
## [107] "Somerset Academy" "George Military College"   
## [109] "Manhattan Middle School" "Three Oaks Middle School"   
## [111] "Hamilton HS West" "The Howard School"   
## [113] "McLean High School" "Gulf Elementary"   
## [115] "Viera HS" "Naples/Barron HS"   
## [117] "Ashley HS" "Upper Valley"   
## [119] "Plantation Key ES" "Sandhill"   
## [121] "Sarasota" "Unity School"   
## [123] "Morganton Day School" "Riverglades"   
## [125] "Wellington" "Woodland Park MS"   
## [127] "Mt. Paran HS" "St. Stephen's Episcopal"   
## [129] "Concordia JR HS" "Two Rivers HS"   
## [131] "Gerstell Academy" "AL Insitute for Deaf & Blind"  
## [133] "Ocean Reef Camp" "Teacher Workshop"   
## [135] "Teacher Workshop-Advanced" "Teacher Workshop-CS"

Lots of synonyms in this list so we will recode those groups that are entered differently but the same.

group\_recode <- instructor\_recode %>% mutate(group\_name = recode(group\_name,   
 "Hamilton West HS" = "Hamilton HS West",  
 "Hamilton West" = "Hamilton HS West",  
 "Hamilton High" = "Hamilton HS West",  
 "New Hanover HS" = "New Hanover",  
 "Episcopal" = "St. Stephen's Episcopal",  
 "St. Paul Lutheren" = "St. Paul",  
 "Naples/Barron HS" = "Barron/Naples",  
 "Naples / Barron" = "Barron/Naples",  
 "Naples" = "Barron/Naples",  
 "Georgia Military" = "GMC",  
 "Georgia Military Academy" = "GMC",  
 "GMC HS" = "GMC",  
 "Woodland" = "Woodland Park",  
 "Woodland Park MS" = "Woodland Park",  
 "Two Rivers HS" = "Two Rivers",  
 "Fairfield" = "Fairfield Middle",  
 "Lutheren" = "Lutheren Schools MW",  
 "Classen SAS" = "Classen",  
 "Dive Ia" = "Dive In",  
 "Teacher Workshop" = "TWS",  
 "Teacher Workshop-Advanced" = "TWS",  
 "Teacher Workshop-CS" = "TWS",  
 "TWS Advanced" = "TWS",  
 "Treasure Village Montessori" = "Treasure Village",  
 "Plantation Key" = "Plantation Key School",  
 "Plantation Key ES" = "Plantation Key School",  
 "Monarch HS" = "Monarch High School",  
 "Viera" = "Viera HS",  
 "Vierra" = "Viera HS",  
 "Paran" = "Mount Paran",  
 "Mt. Paran HS" = "Mount Paran",  
 "Manitowoc" = "UM Manitowoc",  
 "McLean HS" = "McLean High School",  
 "McClean" = "McLean High School",  
 "Three Oaks Middle School" = "Three Oaks",  
 "Three Oaks MS" = "Three Oaks",  
 "Bell" = "BELL",  
 "Key Largo School" = "KLS",  
 "Ocean Reef" = "ORC",  
 "Marinelab" = "MarineLab",  
 "MarineLab Phyto" = "MarineLab",  
 "Alabama school for the deaf" = "AL Insitute for Deaf & Blind",  
 "Lutz" = "Lutz Prep",  
 "Unity" = "Unity School"))

Lastly, we will need to break the date into three new variables Month, Day, Year for seasonal analysis.

split\_dates <- separate(group\_recode, "numeric\_date", into = c("Month", "Day", "Year"), sep = "\\.", extra = "drop")

Finally, rename the data as clean then take a final glimpse.

WQ\_clean\_data <- split\_dates  
  
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...

Save the data as an RDS.

saveRDS(WQ\_clean\_data, file = "../../data/processed\_data/processeddata.rds")