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)

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"))))))))))))))))))))))))))))))))))))))))