Exploratory Data Analysis

William Norfolk

10/9/2019

Load the required libraries for exploratory analysis.

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(plotly)

##   
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':  
##   
## last\_plot

## The following object is masked from 'package:stats':  
##   
## filter

## The following object is masked from 'package:graphics':  
##   
## layout

library(knitr)  
library(naniar)  
library(broom)

Load the processed data from the RDS. Then take a look!

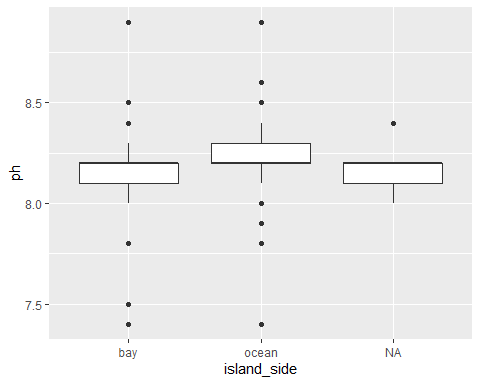
WQ\_clean\_data <- readRDS("../../data/processed\_data/processeddata.rds")  
  
glimpse(WQ\_clean\_data)

## 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...

Now lets generate some plots to compare ocean vs bay.

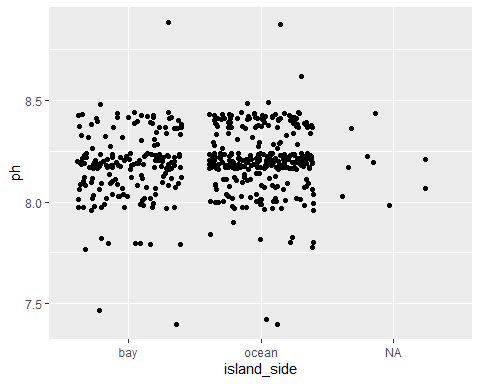
Looks like pH is pretty consistent between the ocean and bayside.

WQ\_clean\_data %>% ggplot() +   
 geom\_boxplot(aes(x = island\_side, y = ph), na.rm = TRUE)



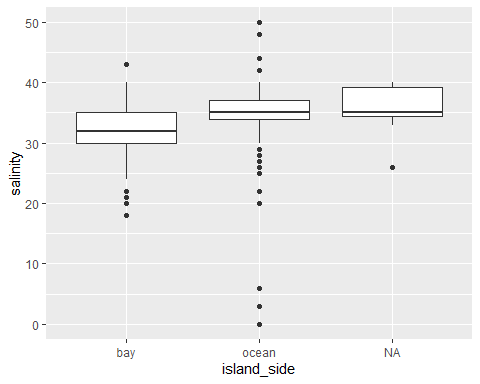
WQ\_clean\_data %>% ggplot() +   
 geom\_jitter(aes(x = island\_side, y = ph))

## Warning: Removed 6 rows containing missing values (geom\_point).



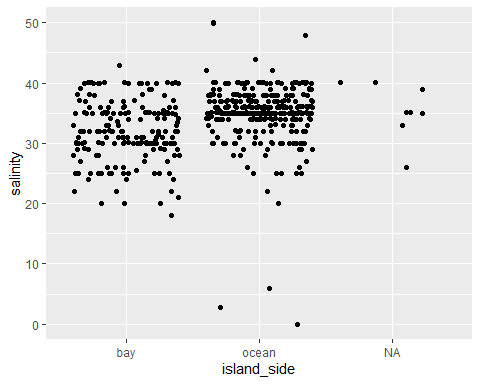
Looks like salinity may be a little more interesting. Ocean side appears to have a narrower range of values compared to the bayside, which is expected due to the size of the body of water.

WQ\_clean\_data %>% ggplot() +   
 geom\_boxplot(aes(x = island\_side, y = salinity), na.rm = TRUE)

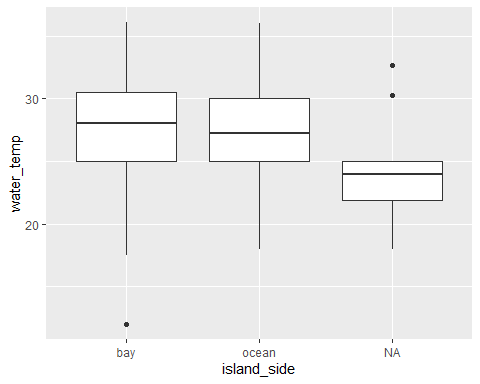


WQ\_clean\_data %>% ggplot() +   
 geom\_jitter(aes(x = island\_side, y = salinity))

## Warning: Removed 16 rows containing missing values (geom\_point).

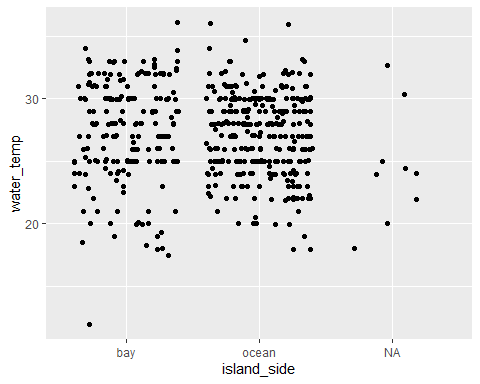


WQ\_clean\_data %>% ggplot() +   
 geom\_boxplot(aes(x = island\_side, y = water\_temp), na.rm = TRUE)

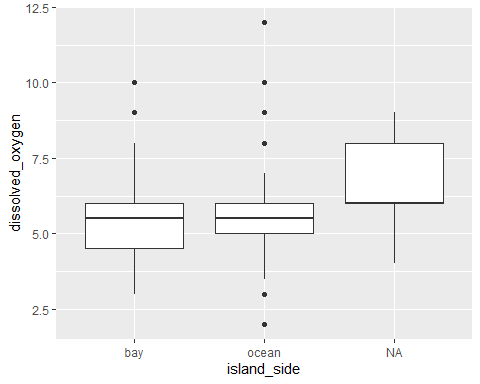


WQ\_clean\_data %>% ggplot() +   
 geom\_jitter(aes(x = island\_side, y = water\_temp))

## Warning: Removed 12 rows containing missing values (geom\_point).

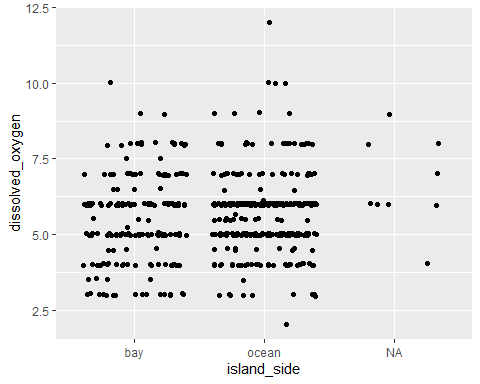


WQ\_clean\_data %>% ggplot() +   
 geom\_boxplot(aes(x = island\_side, y = dissolved\_oxygen), na.rm = TRUE)

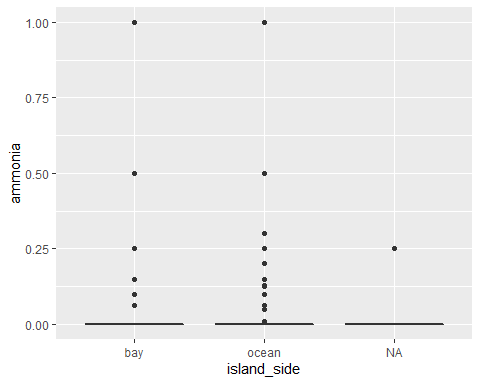


WQ\_clean\_data %>% ggplot() +   
 geom\_jitter(aes(x = island\_side, y = dissolved\_oxygen))

## Warning: Removed 4 rows containing missing values (geom\_point).

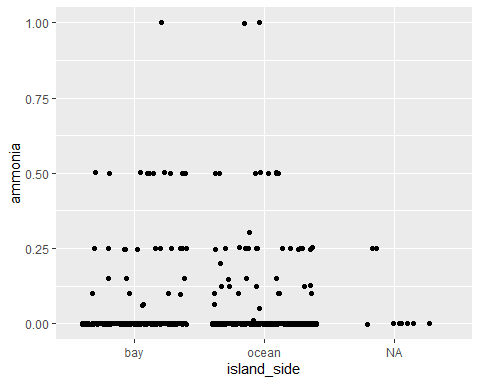


WQ\_clean\_data %>% ggplot() +   
 geom\_boxplot(aes(x = island\_side, y = ammonia), na.rm = TRUE)



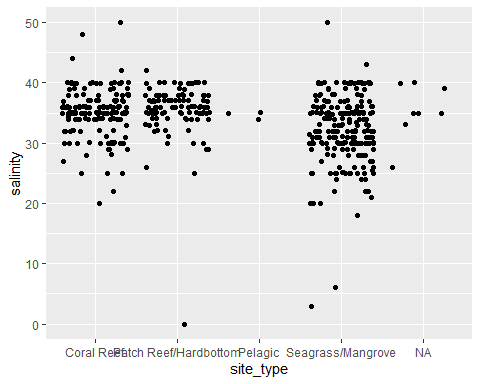
WQ\_clean\_data %>% ggplot() +   
 geom\_jitter(aes(x = island\_side, y = ammonia))

## Warning: Removed 10 rows containing missing values (geom\_point).



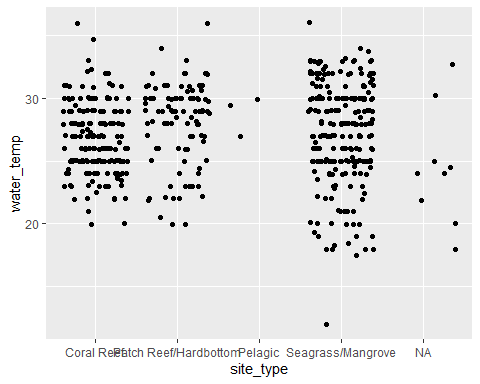
WQ\_clean\_data %>% ggplot() +   
 geom\_jitter(aes(x = site\_type, y = salinity))

## Warning: Removed 16 rows containing missing values (geom\_point).



WQ\_clean\_data %>% ggplot() +   
 geom\_jitter(aes(x = site\_type, y = water\_temp))

## Warning: Removed 12 rows containing missing values (geom\_point).



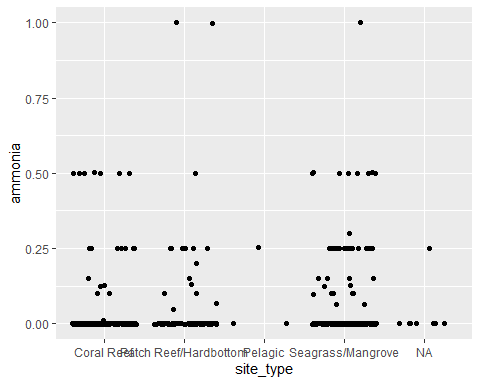
WQ\_clean\_data %>% ggplot() +   
 geom\_jitter(aes(x = site\_type, y = dissolved\_oxygen))

## Warning: Removed 4 rows containing missing values (geom\_point).



WQ\_clean\_data %>% ggplot() +   
 geom\_jitter(aes(x = site\_type, y = ammonia))

## Warning: Removed 10 rows containing missing values (geom\_point).



To Do: 1. Make Bay vs Ocean plots for each paremeter 2. Make a plot of site type for each parameter 3. geom\_line for view hurricane Irma