The Effect of Water Quality Parameters on Dissolved Oxygen Concentrations in Oahu

https://github.com/gdg12/FinalDataAnalyticsProject

Gaby Garcia

Abstract

Water bodies with higher concentrations of dissolved oxygen tend to have more diverse and stable aquatic ecossytems. This data analysis focuses on the water quality parameters that have statistical significance on dissolved oxygen concentrations at coastal monitoring locations on the island of Oahu. A multiple linear regression was performed encompassing all continuous water quality parameters that were not multicollineated with dissolved oxygen. The analysis demonstrates that enterococcus, temperature, pH, and salinity are statistically significant explanatory variables for dissolved oxygen concentrations in Oahu. A spatial analysis of dissolved oxygen concentrations was performed for the North, South, East, and West coasts of Oahu; however, no trend was apparent. A temporal analysis of dissolved oxygen was performed; dissolved oxygen concentrations were lowest in the summer months and highest in the winter months.

Contents

List of Tables

List of Figures

tbls= #Setup

```
setwd("~/Desktop/Environmental Data Analytics/Environmental_Data_Analytics/Final Project
OahuData<-read.csv('OahuData.csv')
```

6.15 Load Necessary Packages

```
library(tidyverse)
library(tidyr)
library(ggplot2)
library(GGally)
library(dplyr)
library(plyr)
library(lubridate)
library(viridis)
library(RColorBrewer)
library(colormap)
library(gridExtra)
library(corrplot)
library(nlme)
library(lsmeans)
library(multcompView)
library(trend)
library(mapview)
library(leaflet)
library(sf)
library(car)
library(stats)
library(wesanderson)
library(scales)
library(extrafont)
```

6.16 Set GGPlot Theme

```
gabytheme <- theme_bw(base_size = 14) +
  theme(plot.title=element_text(face="bold", size="15", color="Indianred4", hjust=0.5),
        axis.title=element_text(face="bold.italic", size=11, color="black"),
axis.text = element_text(face="bold", size=10, color = "black"),
panel.background=element_rect(fill="white", color="darkblue"),
panel.border = element_rect(color = "black", size = 2),
legend.position = "top", legend.background = element_rect(fill="white", color="black"),
```

```
            legend.key = element_rect(fill="transparent", color="NA"))
theme_set(gabytheme)
```

7 Research Question and Rationale

The economic well being of island states such as Hawaii partially depends on thriving coastal ecosystems. Therefore, it is important to analyze time series water quality data featuring parameters such as dissolved oxygen. Dissolved oxygen (DO) is one of the best indicators of a water body's health. Water bodies with higher DO concentrations are healthier and support diverse aquatic plants and animals. Water bodies with low DO concentrations are often severely polluted. The amount of dissolved oxygen that a specific water body can hold is a function of altitude (atmospheric pressure), volume of water, amount of nutrients in the water, aquatic organisms in the water, salinity, water temperature, and the amount of other substances dissolved in the water, in addition to more parameters. Cool water can hold more oxygen than warm water, with variations ranging from seasonal to time of day or night. Water with higher salinity has a lower DO concentration than water with lower salinity at the same temperature. I am interested in examining what water quality parameters affects dissolved oxygen concentrations in Hawaii.

After some searching for water quality data on data.gov, I found a dataset from the Monitoring Section of the State of Hawaii, Department of Health, Clean Water Branch (CWB). The CWB collected water quality field parameters at over 300 coastal monitoring locations in all of Hawaii's islands using well-established instruments and methodologies. The original dataset I found on data.gov spanned 1999-2006. Because I wanted to account for more recent data, I found the complete dataset spanning through 2019 on the Hawaii Clean Water Branch's website. The complete list of relevant parameters are: Location Identifier, Location Name, Island, Latitude of Sample, Longitude of Sample, Sample Number, Date Sample was Collected, Time Sample was collected, Enterococci, Temperature, Salinity, Dissolved Oxygen, Percent Saturation of Dissolved Oxygen, pH, turbidity, Specific Heat of Water (Cp), and relevant comments about water/weather conditions (if applicable). I decided to filter the data for the island of Oahu only because the original dataset included data for all of the islands, and Oahu is home to two-thirds of Hawaii's population and hosts Honolulu, the capital. My reasoning was that a higher island population could lead to more water body pollution. In addition, I left out Percent Saturation of Dissolved Saturation as an explanatory variable because

The null hypothesis of this statistical test is that pH, turbidity, salinity, temperature, enterococcus concentrations, and specific heat have no effect on the dissolved oxygen concentrations in Oahu. This would be represented by the equation xYi=Bo+Ei. The alternative hypothesis is that at least one of the predictor variables have an effect on dissolved oxygen concentrations in Oahu, which would be represented by the equation xYi=Bo +B1X1 +BnXn+... +Ei. My research questions are: 1.) Which of the parameters have a statistically significant relationship with dissolved oxygen concentrations? Of these relevant parameters, which of

them have the most significant effect on dissolved oxygen concentrations over time? What is the magnitude and direction of any effects? 2. Do dissolved oxygen concentrations vary spatially across the island of Oahu?

8 Dataset Information

Water Quality Data was sampled by the State of Hawaii Department of Health Clean Water Branch (specifically, their Beach Monitoring Program).The dataset contains data from the Statewide Water Quality Sampling Dataset from 373 coastal monitoring sites in Hawaii from 1999-2019. All sampling and testing was done in accordance with test procedures approved under 40 CFR Part 136 unless other test procedures have been specified in the permit or approved by the director.

Data only from the Island of Oahu was selected from 01/01/1999 to 04/12/2019.

8.1 Omit NA's from Data (GLM 12 lesson says to do so)

```
OahuDataClean<- na.omit(OahuData)
```

8.2 Change Date from Factor to Date Object

```
OahuDataClean$Date<-as.Date(OahuDataClean$Date, format = "%m/%d/%y")
```

8.2.1 Add a Week, Month, and Year Column to Dataframe Using Mutate Function

```
OahuDataClean<-mutate(OahuDataClean, Week = week(Date))
OahuDataClean<- mutate(OahuDataClean, Month = month(Date))
OahuDataClean<- mutate(OahuDataClean, Year = year(Date))
```

8.3 Create Data Structure Table

```
OahuDataCleanSummary<-OahuDataClean%>%
  group_by(Month)%>%
 summarize(MeanDO=mean(DO, na.rm=TRUE), MeanDOPercentSaturation=mean(DissolvedOxygenSatu
```

9 Exploratory Data Analysis and Wrangling

They only want to see code for data exploration and wrangling, not for data visualization Make sure to write a paragraph discussing whatI did to wrangle my data

• Your data exploration section should contain at least five lines of code that generate summary information about your dataset (or components therein) • Your data exploration section should contain at least three graphs • Your statistical modeling section should contain at least three tests, including the rationale why you took the approach you did. Ensure you have met assumptions of tests or that you justify moving forward without meeting assumptions. • Your data visualization section should contain at least three graphs.

9.1 Structure of Water Data

```
str(OahuDataClean)
```

```
## 'data.frame':    22991 obs. of  22 variables:
##  $ LocationIdentifier       : int  177 253 263 289 265 171 249 273 233 250 ...
##  $ LocationName             : Factor w/ 166 levels "Ala Moana Lagoon",..: 134 85 150
##  $ Island                   : Factor w/ 1 level "Oahu": 1 1 1 1 1 1 1 1 1 1 ...
##  $ LatDecDeg                : num  21.6 21.6 21.7 21.7 21.6 ...
##  $ LongDecDeg               : num  -158 -158 -158 -158 -158 ...
##  $ CP                       : Factor w/ 13 levels "","<",">","E",..: 2 1 2 2 2 2 2 2
##  $ CP.Result                : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ Ent                      : Factor w/ 14 levels "","<","<\xca",..: 1 1 1 1 1 1 1 1
##  $ Enterococcus             : num  10 2.3 10 64 2.3 2.3 2.3 2.3 2.3 2.3 ...
##  $ Sample.No                : Factor w/ 23807 levels "AJ01071501","AJ01071502",..: 93
##  $ Date                     : Date, format: "2019-04-09" "2019-04-09" ...
##  $ Time                     : Factor w/ 351 levels "","1:32 PM","10:00 AM",..: 242 27
##  $ Temperature              : num  24.2 24.6 24.5 24.8 24.9 26.4 25.3 25 25.4 25.7 ..
##  $ Salinity                 : num  32 34.7 34.8 35.1 35 ...
##  $ DO                       : num  5.91 6.38 6.44 6.33 6.23 8.35 7.49 6.81 6.56 6.53
##  $ DissolvedOxygenSaturation: num  84.6 93.4 94.1 93.1 91.8 ...
##  $ pH                       : num  8.05 8.11 8.19 8.16 8.15 8.29 8.07 8.06 8.07 8.06
##  $ Turbidity                : num  4.39 4.09 1.05 1.33 1.52 2.19 2.13 2.53 5.28 3.58
##  $ Comments                 : Factor w/ 11294 levels "","(1) kayaker, (1) paddleboard
##  $ Week                     : num  15 15 15 15 15 15 15 15 15 15 ...
##  $ Month                    : num  4 4 4 4 4 4 4 4 4 4 ...
##  $ Year                     : num  2019 2019 2019 2019 2019 ...
##  - attr(*, "na.action")= 'omit' Named int  20 39 52 66 139 143 168 193 194 196 ...
##   ..- attr(*, "names")= chr  "20" "39" "52" "66" ...
```

9.2 Summary of Water Data

```
summary(OahuDataClean)
```

```
##  LocationIdentifier                    LocationName    Island
##  Min.   :152.0      Kahanamoku Beach, Waikiki: 1001   Oahu:22991
##  1st Qu.:176.0      Ala Moana Lagoon         :  994
##  Median :201.0      Ala Moana Park, D.H.     :  978
##  Mean   :205.4      Kuhio Beach, Waikiki     :  965
##  3rd Qu.:225.0      Sans Souci               :  937
##  Max.   :416.0      Hanauma Beach Park       :  893
##                     (Other)                  :17223
##    LatDecDeg       LongDecDeg           CP          CP.Result
##  Min.   : 0.00   Min.   :-158.2          :11534   Min.   :  0.000
##  1st Qu.:21.28   1st Qu.:-158.0   <      :11393   1st Qu.:  1.000
##  Median :21.30   Median :-157.8   >      :   32   Median :  1.000
##  Mean   :21.33   Mean   :-157.6   E      :    7   Mean   :  2.558
##  3rd Qu.:21.44   3rd Qu.:-157.8   EST.   :    7   3rd Qu.:  2.000
##  Max.   :21.71   Max.   :   0.0   EST. > :    6   Max.   :400.000
##                                   (Other):   12
##       Ent         Enterococcus          Sample.No
##         :18832   Min.   :    0.30   AJ01071501:    1
##  \xca<  : 2064   1st Qu.:    2.30   AJ01071502:    1
##  <      : 2009   Median :    2.30   AJ01071503:    1
##  >      :   27   Mean   :   22.63   AJ01071504:    1
##  EST.   :   26   3rd Qu.:   10.00   AJ01071505:    1
##  <\xca  :   21   Max.   :24196.00   AJ01071506:    1
##  (Other):   12                      (Other)   :22985
##       Date                 Time        Temperature       Salinity
##  Min.   :2004-12-27   8:00 AM:  531   Min.   : 0.00   Min.   : 0.00
##  1st Qu.:2007-04-04   8:30 AM:  506   1st Qu.:23.80   1st Qu.:34.86
##  Median :2009-02-25   7:40 AM:  490   Median :24.90   Median :35.20
##  Mean   :2010-08-07   7:30 AM:  470   Mean   :24.81   Mean   :34.75
##  3rd Qu.:2014-07-23   8:10 AM:  466   3rd Qu.:25.93   3rd Qu.:35.48
##  Max.   :2019-04-09   7:45 AM:  457   Max.   :35.27   Max.   :38.13
##                      (Other):20071
##        DO          DissolvedOxygenSaturation       pH
##  Min.   : 0.000   Min.   :  0.00            Min.   : 0.000
##  1st Qu.: 5.480   1st Qu.: 83.30            1st Qu.: 8.000
##  Median : 5.890   Median : 88.70            Median : 8.080
##  Mean   : 5.723   Mean   : 85.74            Mean   : 7.962
##  3rd Qu.: 6.210   3rd Qu.: 92.50            3rd Qu.: 8.150
##  Max.   :91.400   Max.   :155.50            Max.   :78.900
##
```

```
##    Turbidity
##  Min.   :  0.000
##  1st Qu.:  1.930
##  Median :  3.580
##  Mean   :  5.704
##  3rd Qu.:  6.860
##  Max.   :315.000
##
##                                                  Comments          Week
##  sunny, light breeze                                :  409   Min.   : 1.00
##  \xcasunny, light breeze, swimmers                  :  394   1st Qu.:13.00
##  sunny, light breeze, swimmers                      :  308   Median :26.00
##  \xcasunny, light breeze                            :  241   Mean   :25.85
##  partly cloudy, lt winds, sm surf, swimmers, beach:  195   3rd Qu.:38.00
##  partly cloudy, mod winds, sm surf, beach walkers,:  148   Max.   :52.00
##  (Other)                                           :21296
##      Month             Year
##  Min.   : 1.000   Min.   :2004
##  1st Qu.: 3.000   1st Qu.:2007
##  Median : 6.000   Median :2009
##  Mean   : 6.352   Mean   :2010
##  3rd Qu.: 9.000   3rd Qu.:2014
##  Max.   :12.000   Max.   :2019
##
```

9.3 Dimensions of Data

```
dim(OahuDataClean)
```

```
## [1] 22991    22
```

9.4 View First 10 Rows of Data Frame

```
head(OahuDataClean, 10)
```

```
##    LocationIdentifier            LocationName Island LatDecDeg LongDecDeg
## 1                 177      Punaluu Beach Park   Oahu  21.57703  -157.8817
## 2                 253               Kokololio   Oahu  21.62517  -157.9197
## 3                 263             Turtle Bay   Oahu  21.70403  -157.9986
## 4                 289               Pipeline   Oahu  21.66447  -158.0525
## 5                 265 Pupukea at Shark's Cove   Oahu  21.64608  -158.0637
## 6                 171      Haleiwa Beach Park   Oahu  21.59840  -158.1036
```

```
## 7                249         Maipalaoa Beach   Oahu  21.40469  -158.1777
## 8                273          Puuohulu Beach   Oahu  21.39647  -158.1637
## 9                233             Nanaikapono   Oahu  21.38639  -158.1514
## 10               250           Ulehawa Beach   Oahu  21.38246  -158.1473
##    CP CP.Result Ent Enterococcus   Sample.No       Date     Time
## 1   <         1             10.0  JL04091901 2019-04-09  8:10 AM
## 2             1              2.3  JL04091902 2019-04-09  8:41 AM
## 3   <         1             10.0  JL04091903 2019-04-09  9:21 AM
## 4   <         1             64.0  JL04091904 2019-04-09  9:51 AM
## 5   <         1              2.3  JL04091905 2019-04-09 10:16 AM
## 6   <         1              2.3  JL04091906 2019-04-09 10:40 AM
## 7   <         1              2.3 MLH04091901 2019-04-09  9:20 AM
## 8   <         1              2.3 MLH04091902 2019-04-09  9:45 AM
## 9             1              2.3 MLH04091903 2019-04-09 10:15 AM
## 10  <         1              2.3 MLH04091904 2019-04-09 10:45 AM
##    Temperature Salinity   DO DissolvedOxygenSaturation   pH Turbidity
## 1         24.2    32.04 5.91                       84.6 8.05      4.39
## 2         24.6    34.73 6.38                       93.4 8.11      4.09
## 3         24.5    34.80 6.44                       94.1 8.19      1.05
## 4         24.8    35.15 6.33                       93.1 8.16      1.33
## 5         24.9    35.01 6.23                       91.8 8.15      1.52
## 6         26.4    34.04 8.35                      125.5 8.29      2.19
## 7         25.3    35.16 7.49                      111.2 8.07      2.13
## 8         25.0    35.11 6.81                      100.7 8.06      2.53
## 9         25.4    34.96 6.56                       97.5 8.07      5.28
## 10        25.7    35.16 6.53                       97.7 8.06      3.58
##                                                          Comments Week
## 1       mostly cloudy; light wind; stream flowing (right); debris   15
## 2                    mostly cloudy; light wind; small shore break   15
## 3  mostly cloudy; light wind; shallow, tidepool area; beach goers   15
## 4          mostly cloudy; light wind; waves; surfers; beach goers   15
## 5                   mostly cloudy; light wind; waves; beach goers   15
## 6        partly cloudy/sunny; shallow, tidepool area; beach goers   15
## 7                                          \xcasunny, light wind   15
## 8                                          \xcasunny, light wind   15
## 9                                          \xcasunny, light wind   15
## 10                              \xcasunny, light wind, wave action   15
##    Month Year
## 1      4 2019
## 2      4 2019
## 3      4 2019
## 4      4 2019
## 5      4 2019
## 6      4 2019
## 7      4 2019
```

```
## 8      4 2019
## 9      4 2019
## 10     4 2019
```

9.5 View all Column Names

```
colnames(OahuDataClean)
```

```
##  [1] "LocationIdentifier"        "LocationName"
##  [3] "Island"                    "LatDecDeg"
##  [5] "LongDecDeg"                "CP"
##  [7] "CP.Result"                 "Ent"
##  [9] "Enterococcus"              "Sample.No"
## [11] "Date"                      "Time"
## [13] "Temperature"               "Salinity"
## [15] "DO"                        "DissolvedOxygenSaturation"
## [17] "pH"                        "Turbidity"
## [19] "Comments"                  "Week"
## [21] "Month"                     "Year"
```

9.6 Omit NA's from Data

```
OahuDataClean<- na.omit(OahuData)
```

9.7 Attach Oahu Data Clean

```
attach(OahuDataClean)
```

10 Recoding Geographical Observations to Regions in Oahu

10.1 Convert Location Identifier from number to factor

```
OahuDataClean$LocationIdentifier<-as.factor(OahuDataClean$LocationIdentifier)
```

10.2 Use Revalue function to recode Station Numbers into Geographical Regions

```
library(plyr)
OahuDataClean$Region<-revalue(OahuDataClean$LocationIdentifier, c('152'='South', '153'='
```

10.3 Remove Observations that are not in Oahu (errors in original dataset)

```
OahuDataClean<-OahuDataClean[-c(18913,19032,19156, 19287, 19384, 19476, 19567, 19745, 19
```

10.4 Remove observations with 0 degrees Long and 0 degrees latitude (incorrect values)

```
library(dplyr)
OahuDataClean<- filter(OahuDataClean, LocationIdentifier != 324 &
                                      LocationIdentifier != 325 &  LocationIdentifier !=
                                      LocationIdentifier != 327 & LocationIdentifier != 3
```

10.5 Determine Number of Observations for Each Region in Oahu

```
summary(OahuDataClean$Region)
```

```
## South North  East  West
## 12262  2638  4694  3316
```

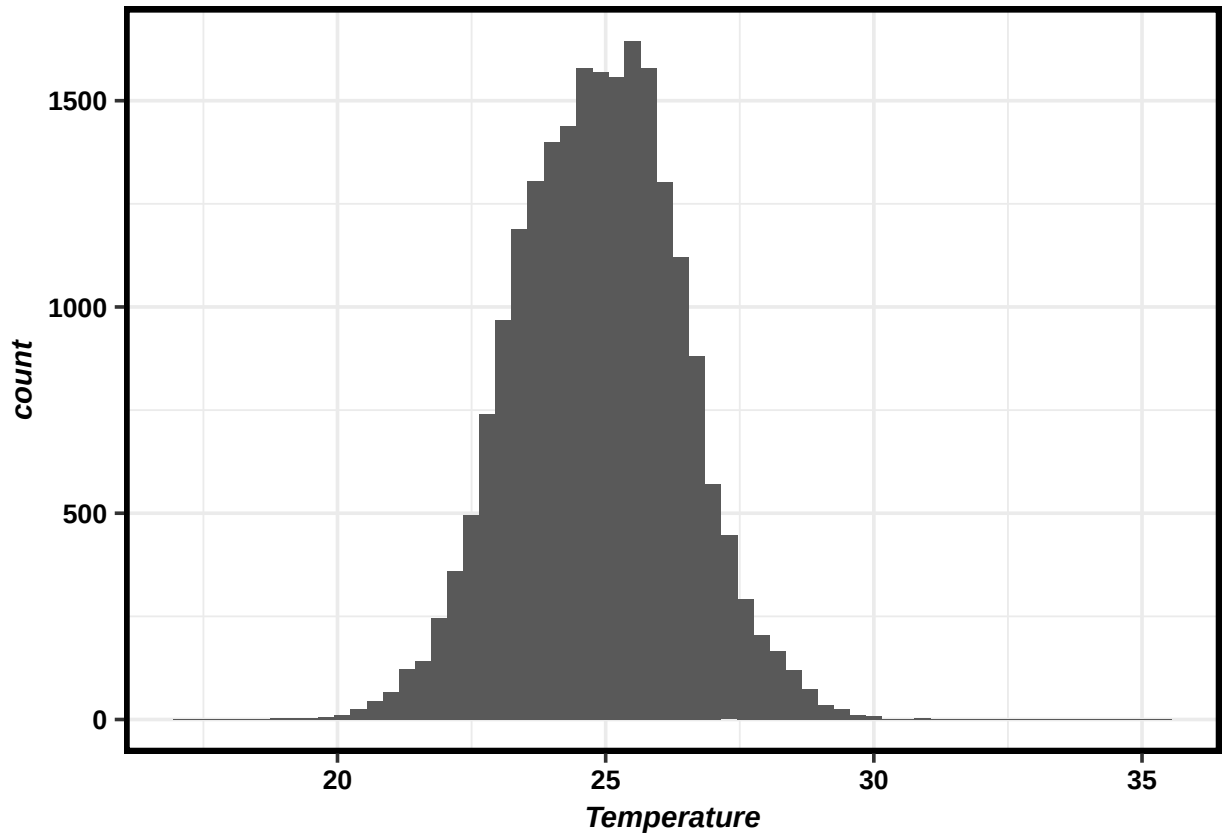
10.6 Remove observations of 0 (can't log-transform data with 0's)

```
 OahuDataClean2<-OahuDataClean %>%   ###pipe indicates then
  filter(Temperature!=0, Salinity!=0, DO!=0, DissolvedOxygenSaturation!=0, pH!=0, Turbid
```

11 Exploration of Continuous Variables

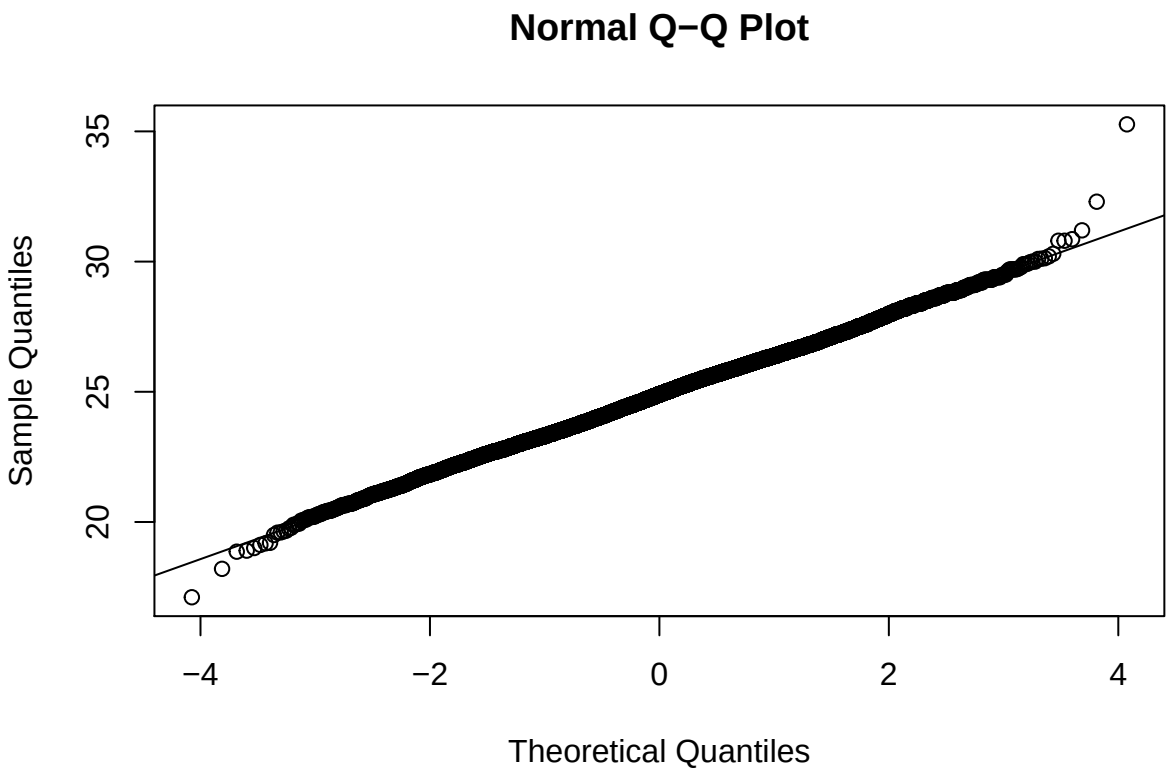
11.1 Temperature

```
ggplot(OahuDataClean2) +
  geom_histogram(aes(x =Temperature), binwidth=0.3)
```



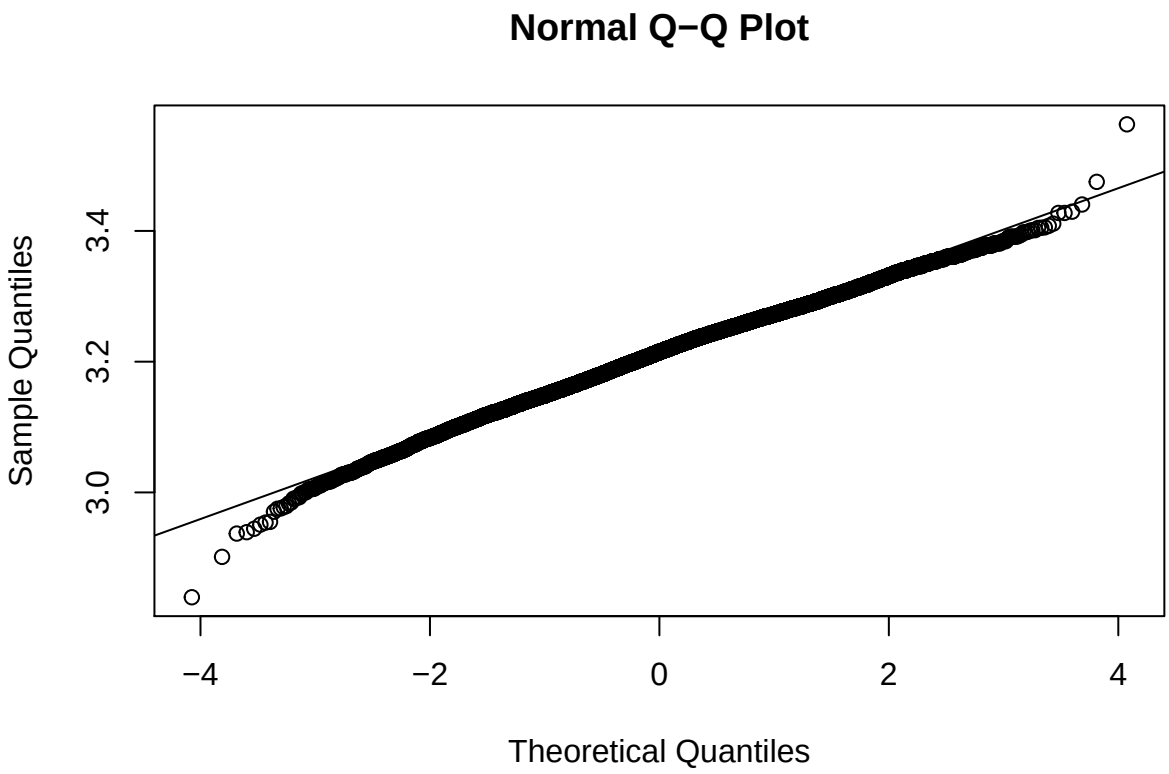
11.2 QQNorm for Temperature

```
 qqnorm(OahuDataClean2$Temperature)
qqline(OahuDataClean2$Temperature)
```



11.3 LogTransform Temperature-doesn't look any better

```
qqnorm(log(OahuDataClean2$Temperature))
qqline(log(OahuDataClean2$Temperature))
```



11.4 Perform Shapiro Wilks Normality test for first 5,000 Temperature observations

```
shapiro.test(OahuDataClean2$Temperature[0:5000])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  OahuDataClean2$Temperature[0:5000]
## W = 0.99317, p-value = 9.529e-15
```

11.4.0.1 Reject null hypothesis that Temperature is normally distributed.

11.4.1 Summary of Temperature

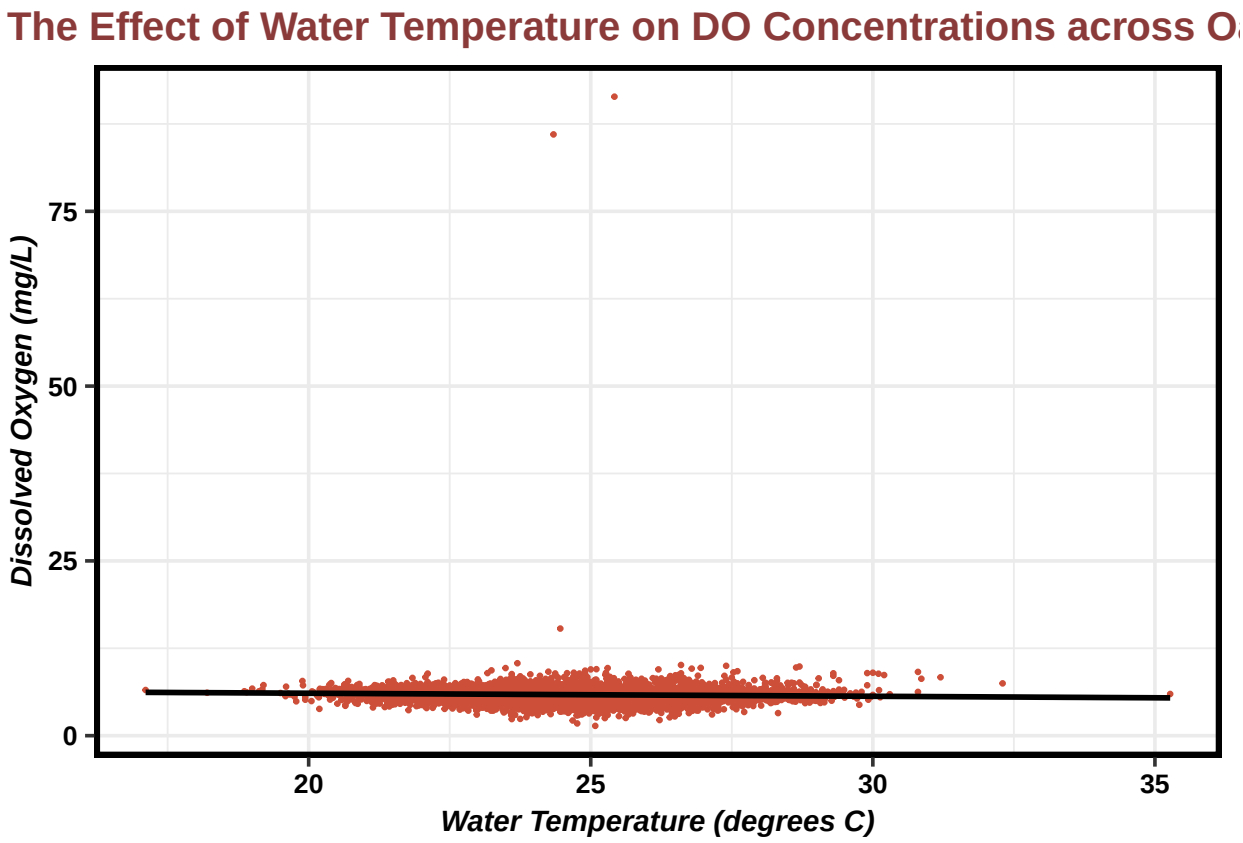
```
summary(OahuDataClean2$Temperature)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##   17.11   23.80   24.90   24.87   25.92   35.27
```

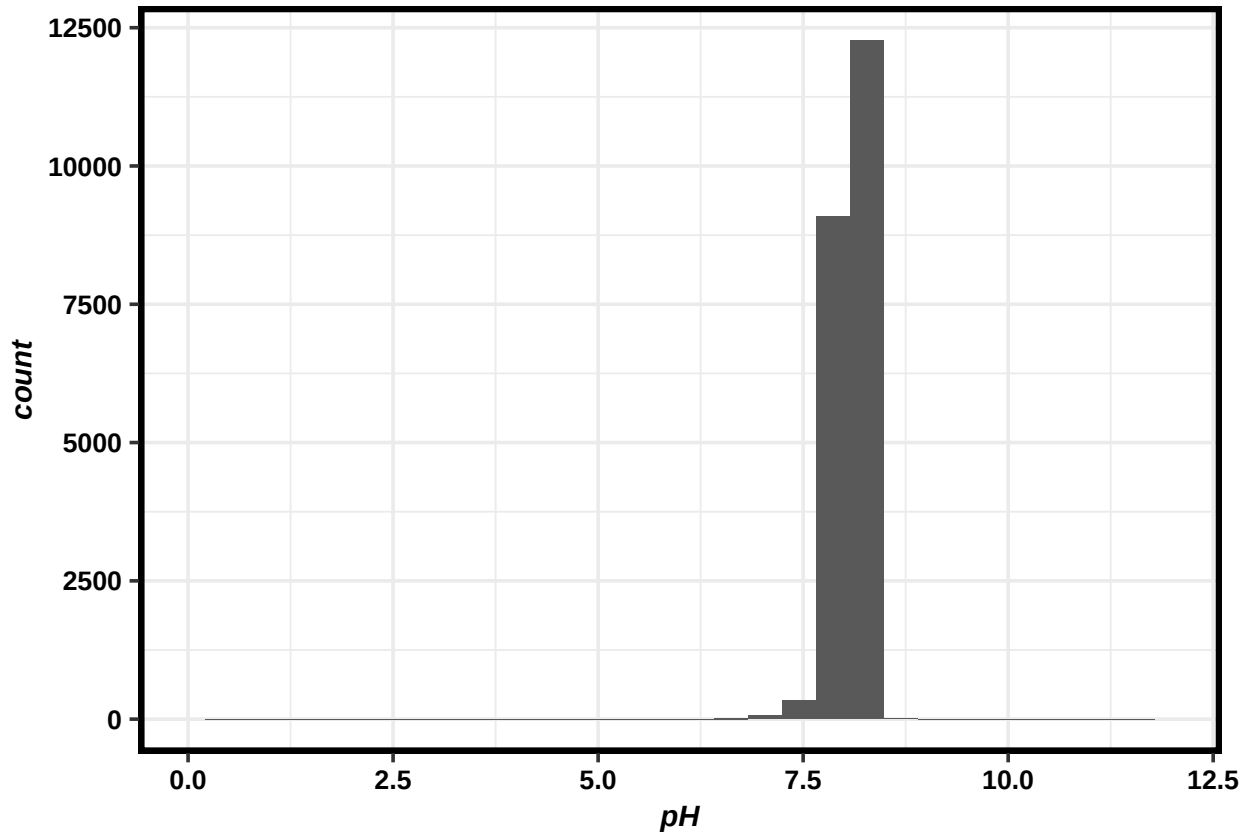
11.5 Plot Temperature against DO

```
TempbyDO <-
  ggplot(OahuDataClean2, aes(x =Temperature, y =DO)) +
  geom_point(color="tomato3", alpha=1, size=0.5) +
  geom_smooth(method=lm, color="black") +
  labs(title="The Effect of Water Temperature on DO Concentrations across Oahu", x="Wate
print(TempbyDO)
```



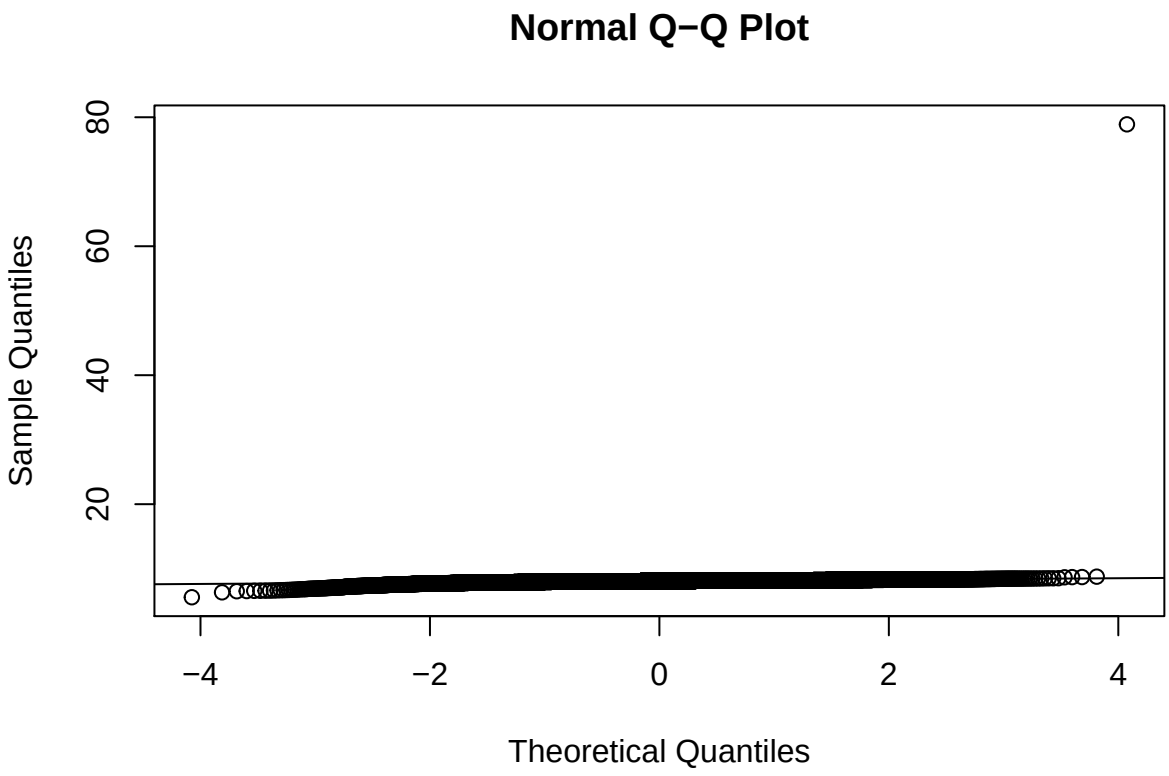
11.6 pH

```
ggplot(OahuDataClean2) +
  geom_histogram(aes(x =pH))+
   scale_x_continuous(limits = c(0, 12))
```



11.7 QQNorm of pH

```
 qqnorm(OahuDataClean2$pH)
qqline(OahuDataClean2$pH)
```



11.8 Shapiro Wilks Test for pH

```
shapiro.test(OahuDataClean2$pH[0:5000])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  OahuDataClean2$pH[0:5000]
## W = 0.70546, p-value < 2.2e-16
```

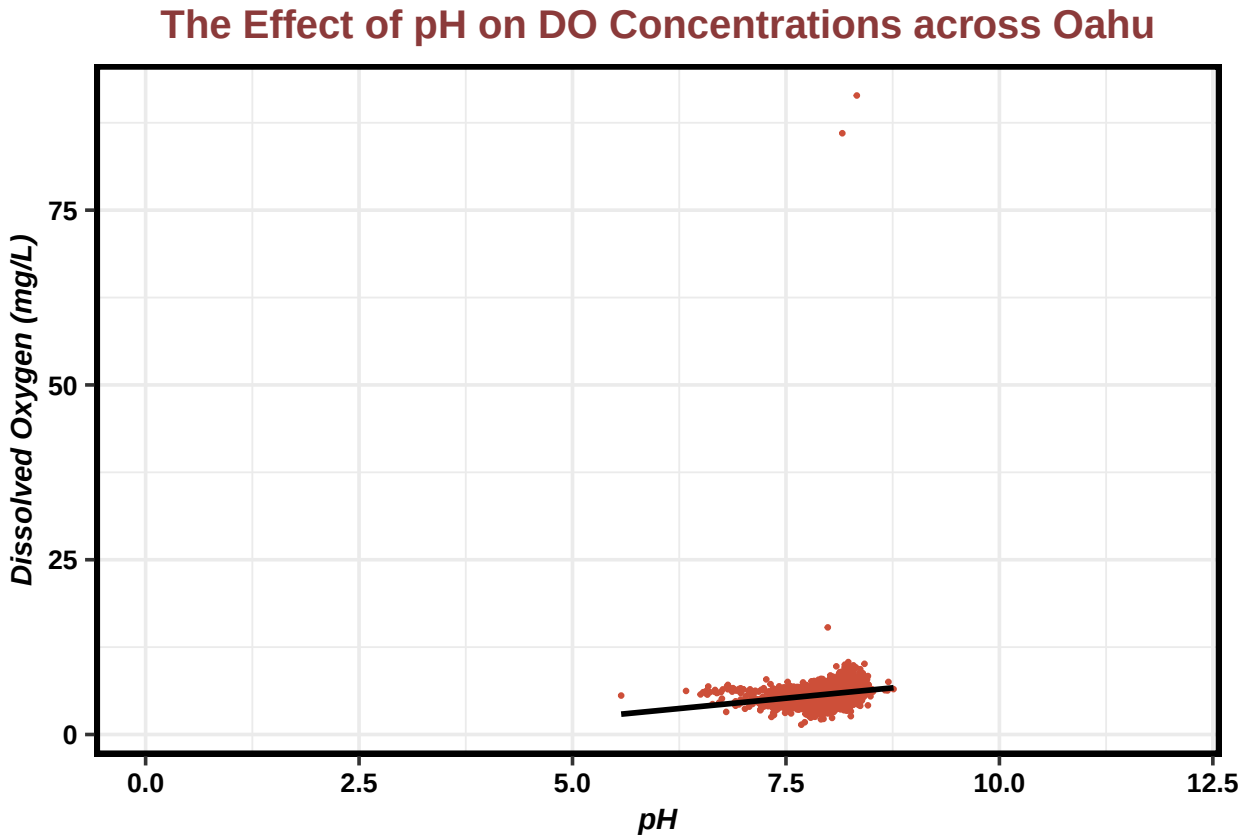
11.8.1 Summary of pH

```
summary(OahuDataClean2$pH)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.570   8.000   8.080   8.064   8.150  78.900
```

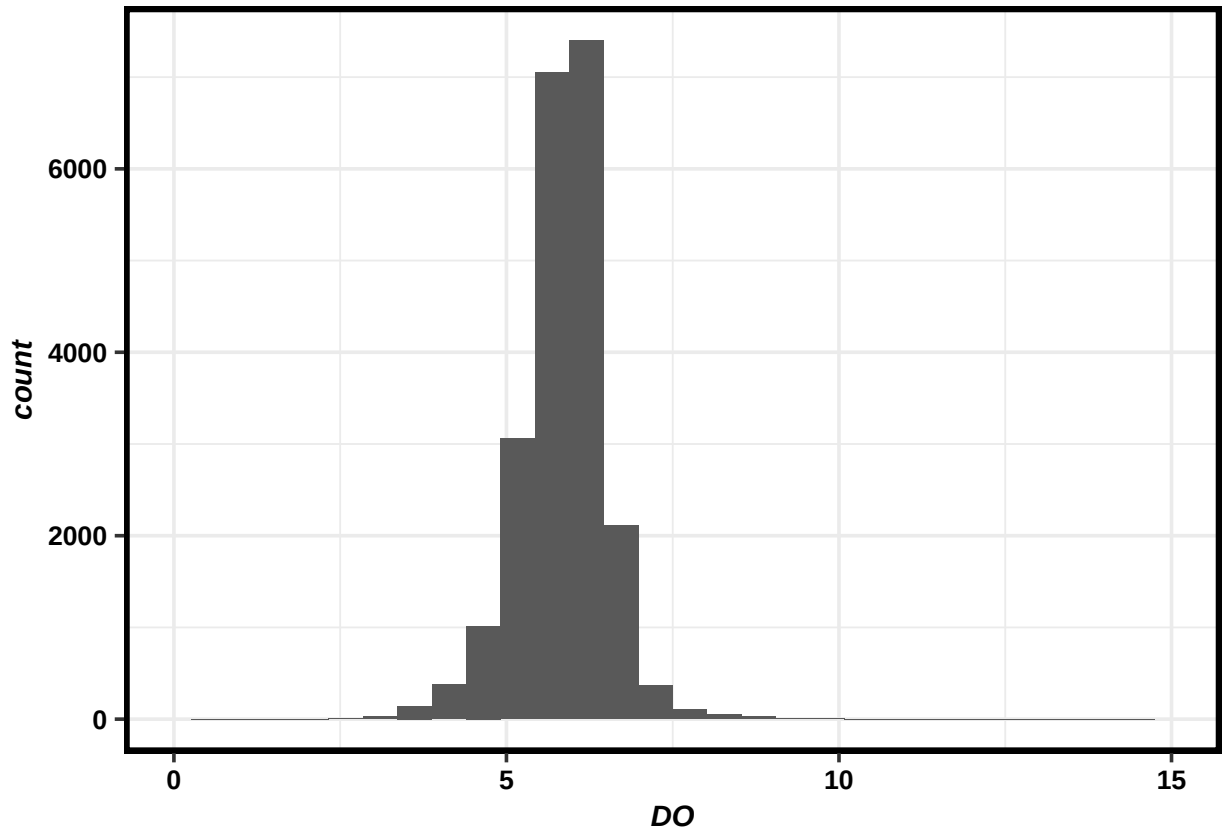
11.8.2 Plot pH against DO

```
pHbyDO <-
  ggplot(OahuDataClean2, aes(x =pH, y =DO)) +
  geom_point(color="tomato3", alpha=1, size=0.5) +
  geom_smooth(method=lm, color="black", se=FALSE) +
  labs(title="The Effect of pH on DO Concentrations across Oahu", x="pH", y="Dissolved O
   scale_x_continuous(limits = c(0, 12))
print(pHbyDO)
```



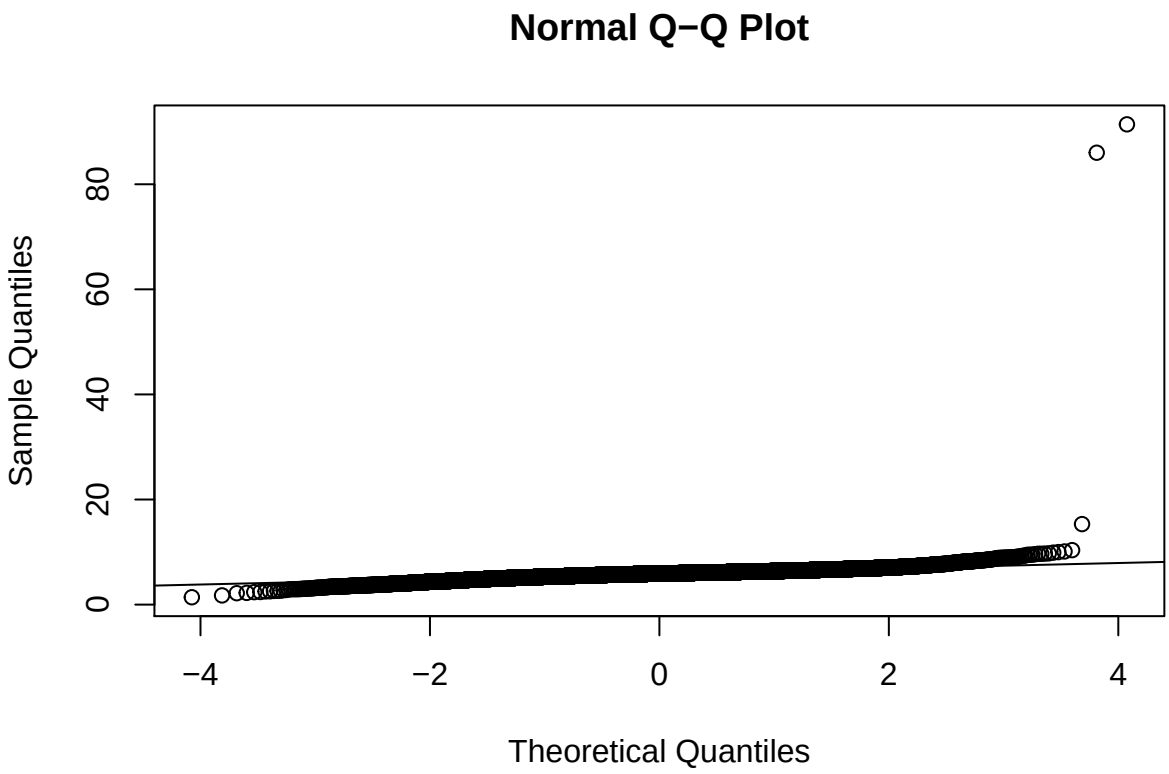
11.9 Dissolved Oxygen Concentrations

```
ggplot(OahuDataClean2) +
  geom_histogram(aes(x =DO)) +
scale_x_continuous(limits = c(0, 15))
```



11.10 QQNorm of DO

```
 qqnorm(OahuDataClean2$DO)
qqline(OahuDataClean2$DO)
```



11.10.1 Shapiro Wilks Test for DO

```
shapiro.test(OahuDataClean2$DO[0:5000])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  OahuDataClean2$DO[0:5000]
## W = 0.9109, p-value < 2.2e-16
```

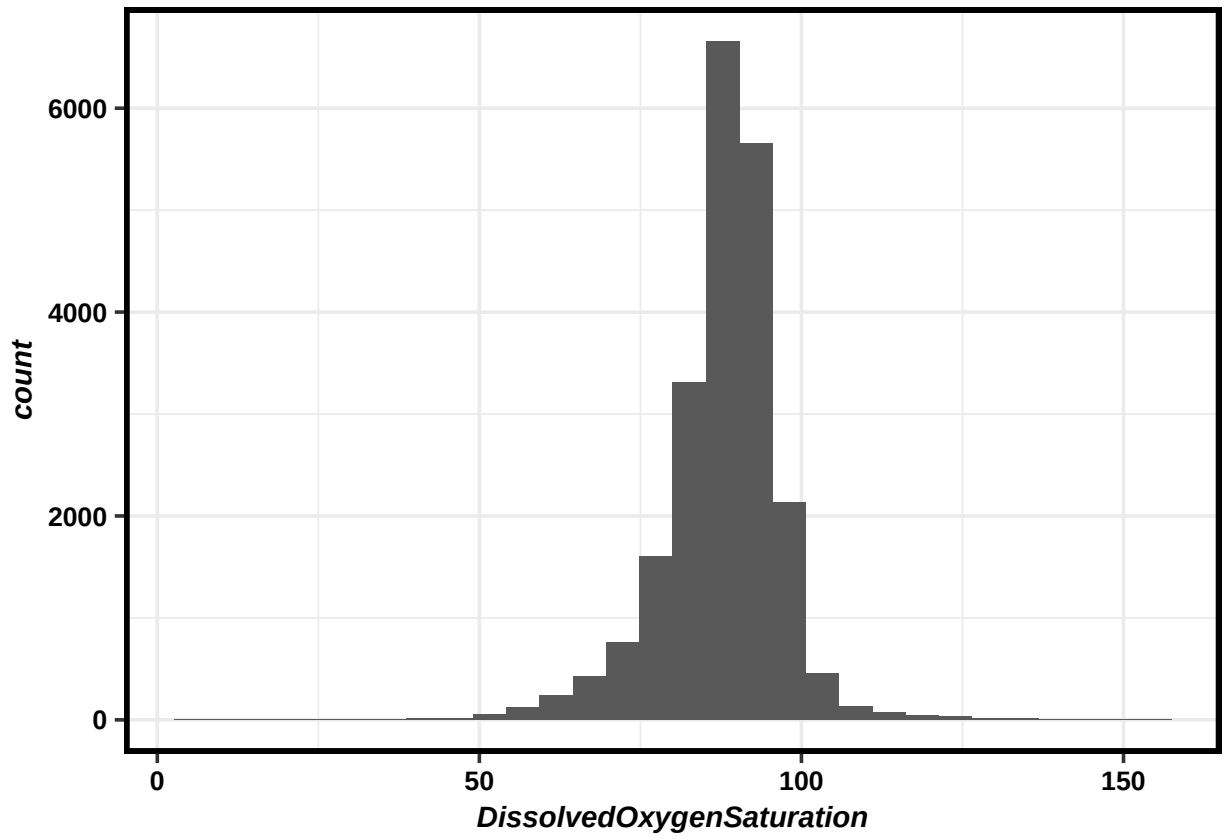
11.11 Summary of DO

```
summary(OahuDataClean2$DO)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.400   5.520   5.900   5.857   6.210  91.400
```

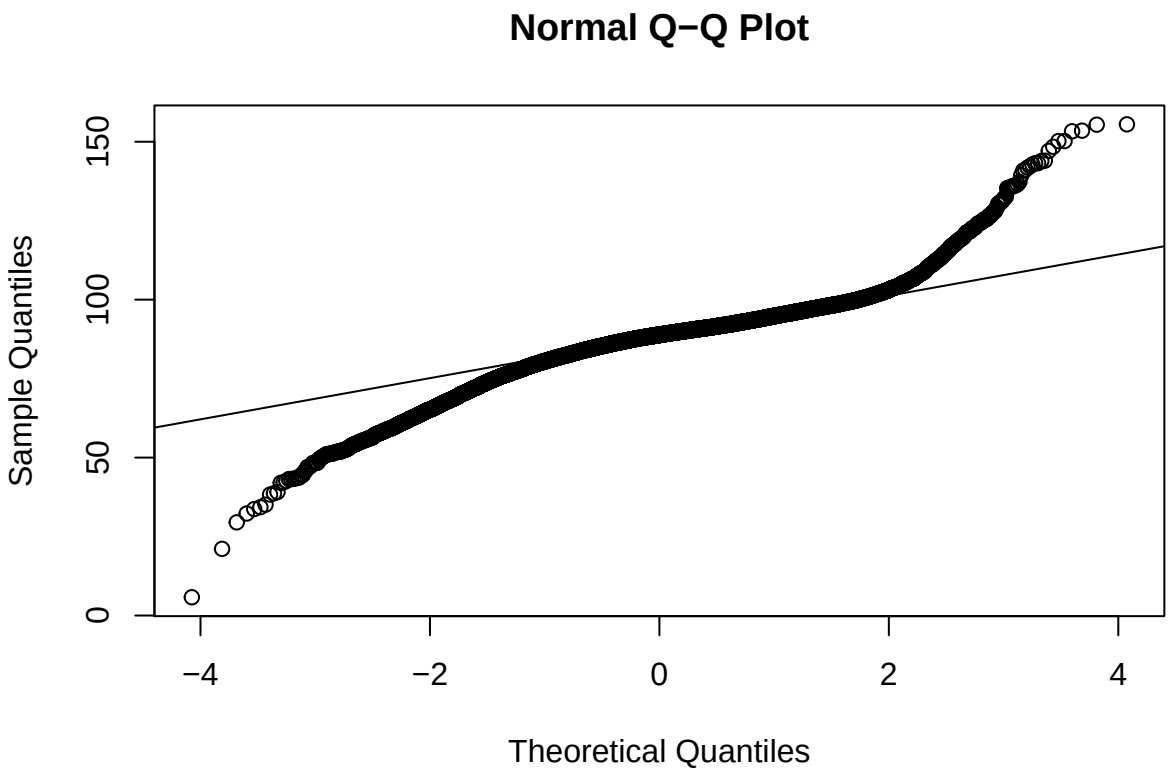
11.12 Percent Saturation of Dissolved Oxygen

```
ggplot(OahuDataClean2) +
  geom_histogram(aes(x =DissolvedOxygenSaturation))
```



11.13 QQnorm of Percent Saturation of Dissolved Oxygen

```
qqnorm(OahuDataClean2$DissolvedOxygenSaturation)
qqline(OahuDataClean2$DissolvedOxygenSaturation)
```



11.14 Shapiro Test for Percent Saturation Dissolved Oxygen

```
shapiro.test(OahuDataClean2$DissolvedOxygenSaturation[0:5000])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  OahuDataClean2$DissolvedOxygenSaturation[0:5000]
## W = 0.87819, p-value < 2.2e-16
```

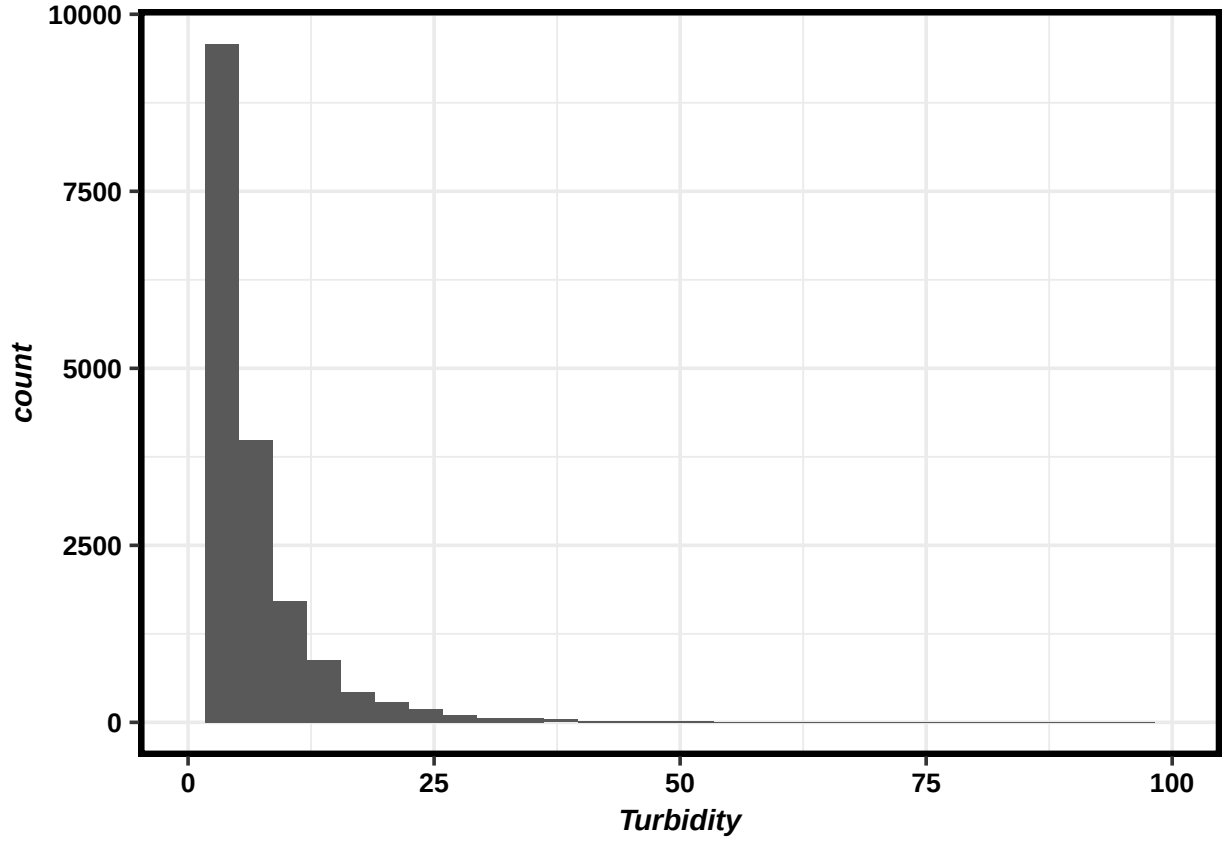
11.14.1 Summary of Percent Saturation of Dissolved Oxygen

```
summary(OahuDataClean2$DissolvedOxygenSaturation)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.79   83.80   88.90   87.76   92.60  155.50
```

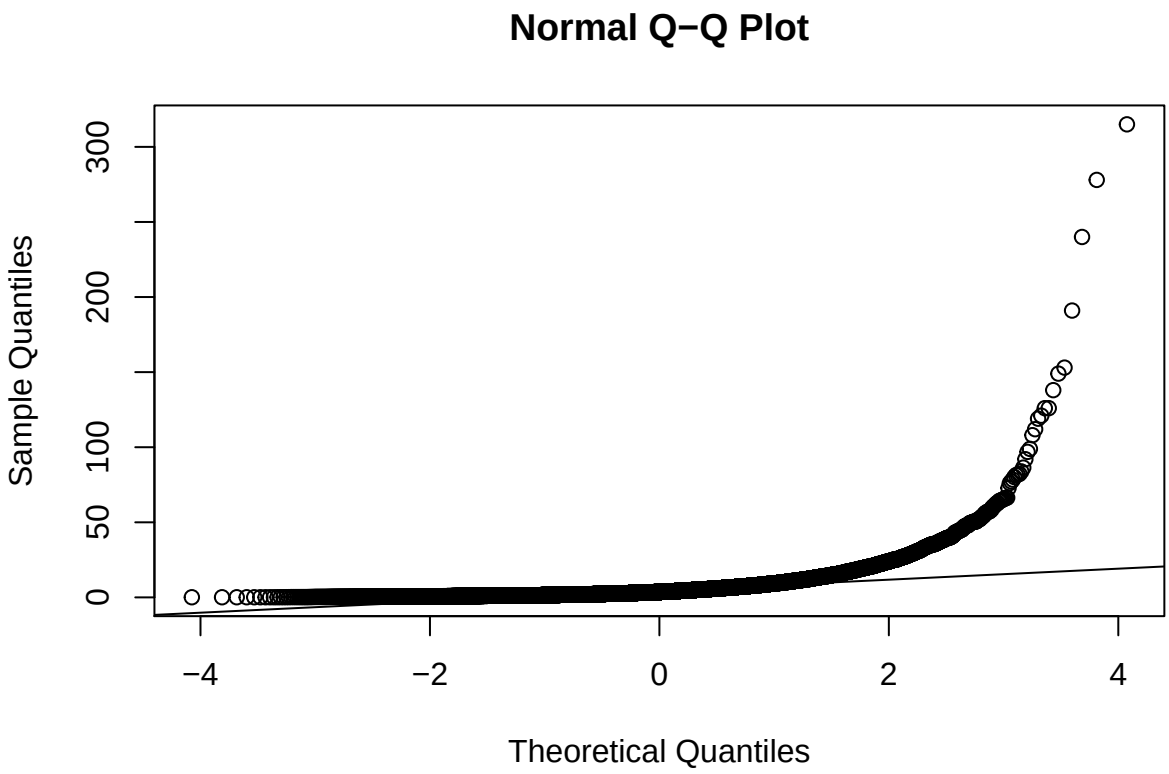
11.15 Turbidity

```
ggplot(OahuDataClean2) +
  geom_histogram(aes(x =Turbidity)) +
   scale_x_continuous(limits = c(0, 100))
```



11.16 QQNorm of Turbidity

```
qqnorm(OahuDataClean2$Turbidity)
qqline(OahuDataClean2$Turbidity)
```



11.16.1 Shapiro Test for Turbidity

```
shapiro.test(OahuDataClean2$Turbidity[0:5000])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  OahuDataClean2$Turbidity[0:5000]
## W = 0.65218, p-value < 2.2e-16
```

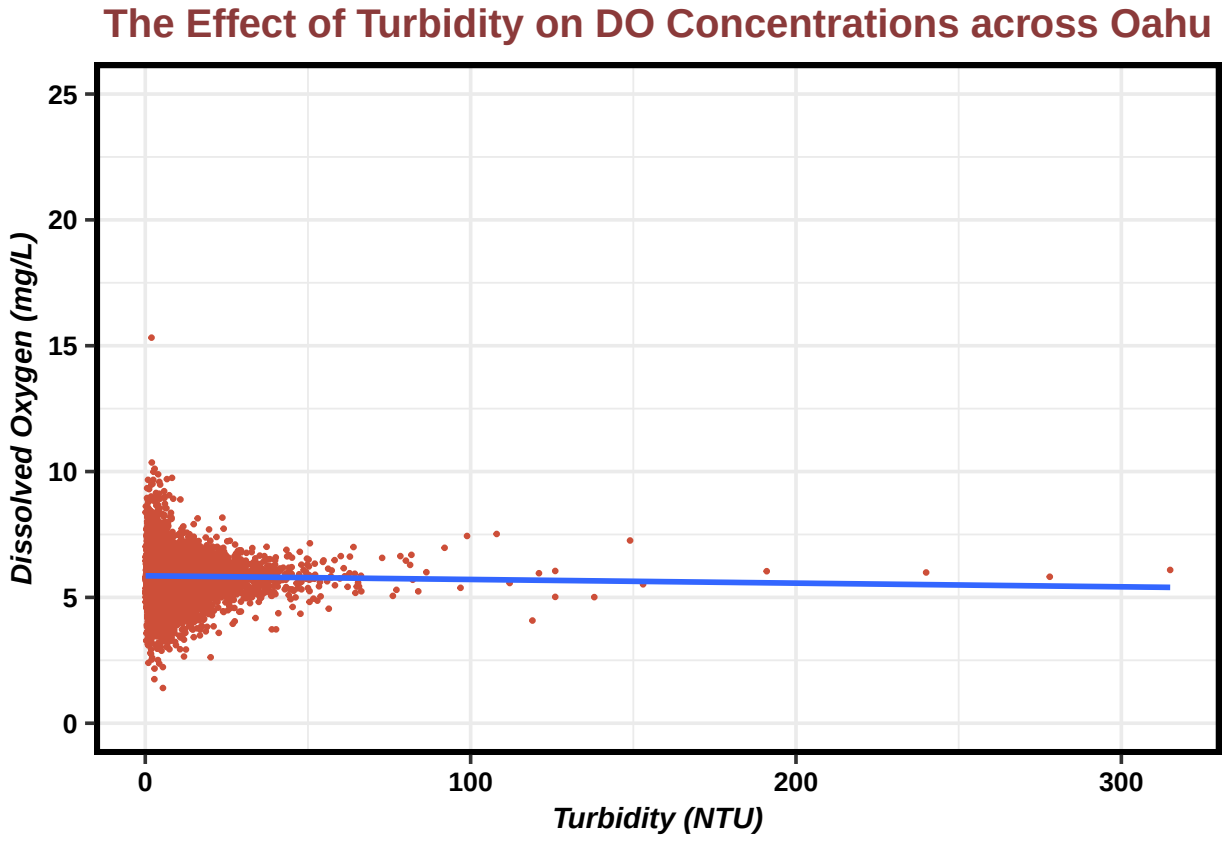
11.16.2 Summary of Turbidity

```
summary(OahuDataClean2$Turbidity)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.100   1.990   3.640   5.779   6.930 315.000
```

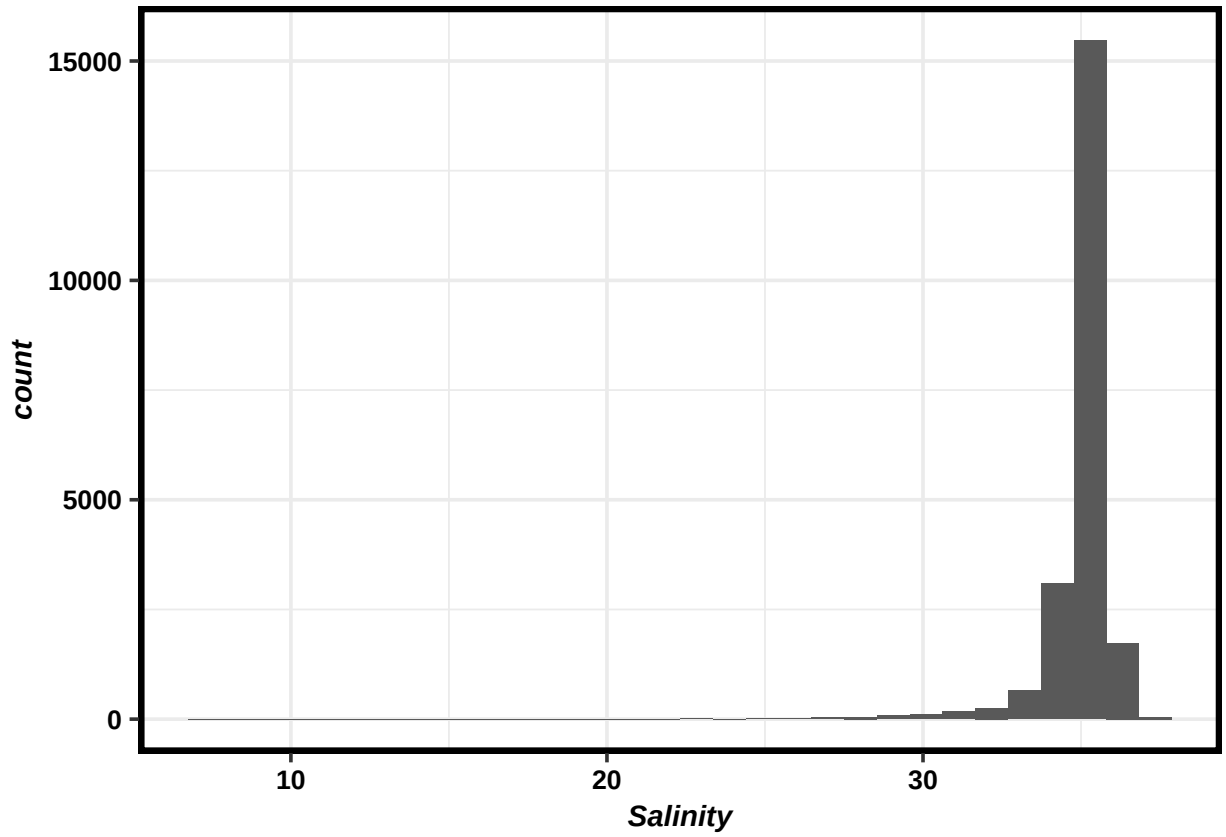
11.17 Plot Turbidity against DO

```
TurbiditybyDO <-
  ggplot(OahuDataClean2, aes(x =Turbidity, y =DO)) +
  geom_point(color="tomato3", alpha=1, size=0.5) +
  geom_smooth(method="lm", se=FALSE) +
   scale_y_continuous(limits = c(0, 25)) +

  labs(title="The Effect of Turbidity on DO Concentrations across Oahu", x="Turbidity (N
print(TurbiditybyDO)
```



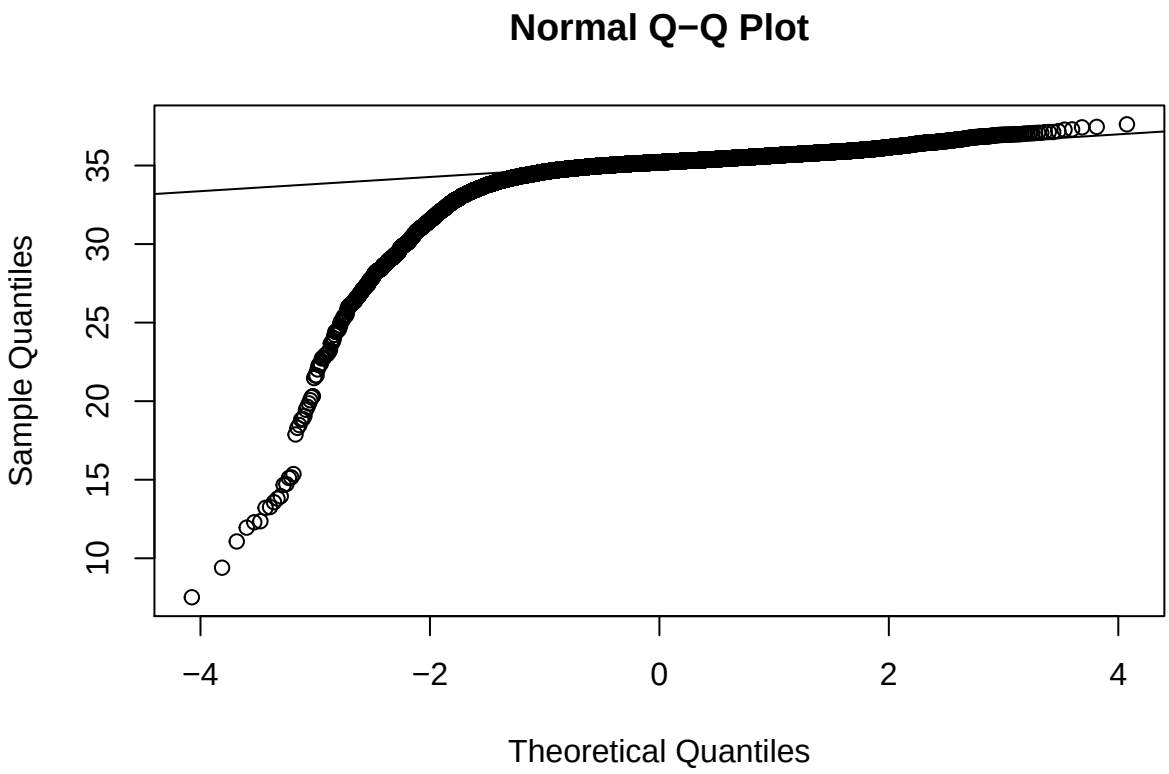
11.18 Salinity

```
ggplot(OahuDataClean2) +
  geom_histogram(aes(x =Salinity))
```



11.19 QQNorm of Salinity

```
qqnorm(OahuDataClean2$Salinity)
qqline(OahuDataClean2$Salinity)
```



11.20 Shapiro Test for Salinity

```
shapiro.test(OahuDataClean2$Salinity[0:5000])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  OahuDataClean2$Salinity[0:5000]
## W = 0.52773, p-value < 2.2e-16
```

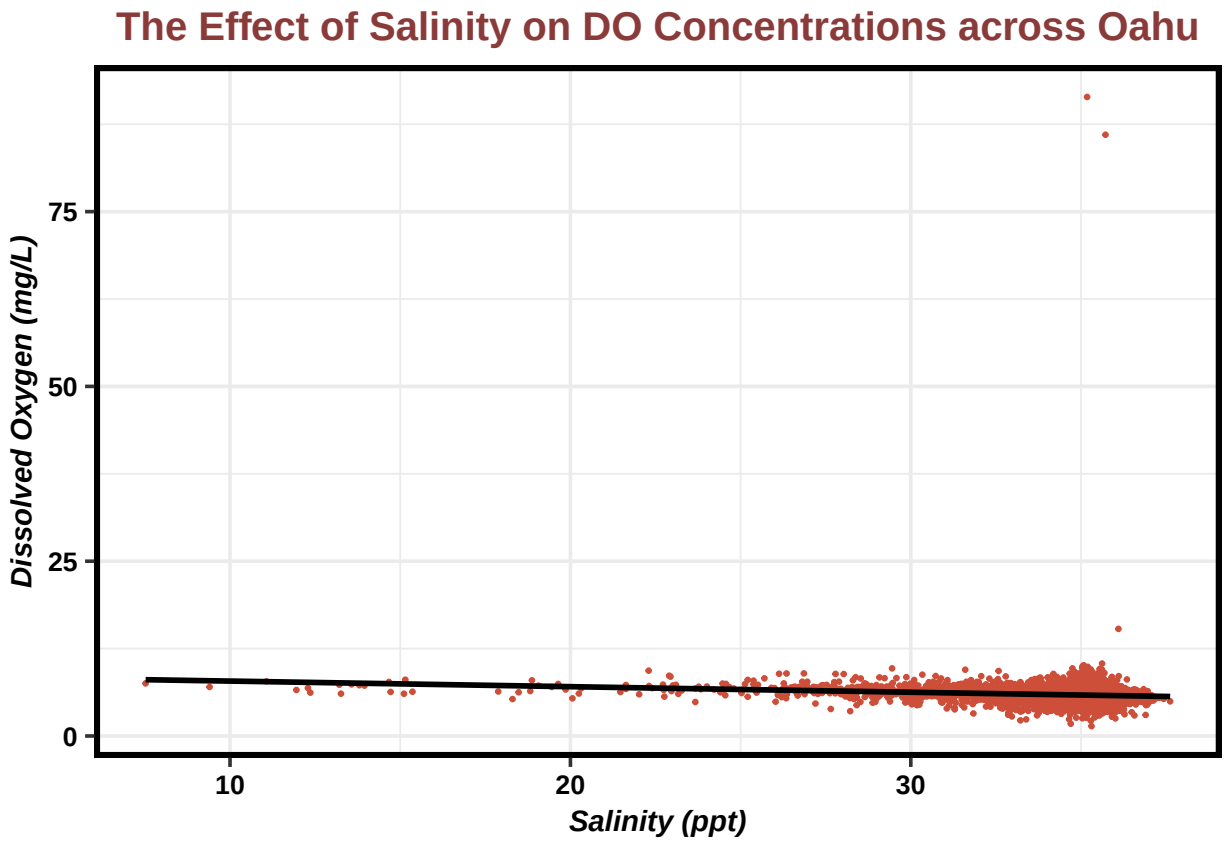
11.20.1 Summary of Salinity

```
summary(OahuDataClean2$Salinity)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.52   34.87   35.20   34.96   35.48   37.62
```

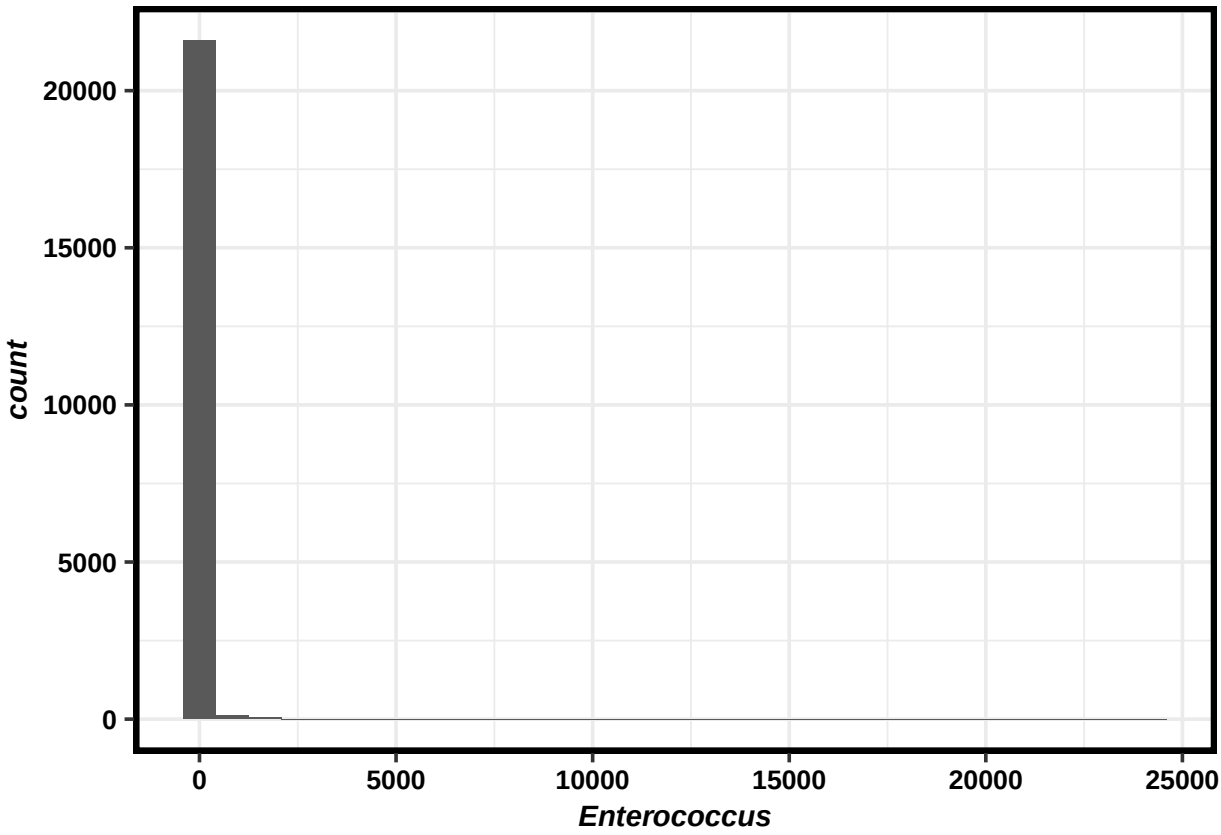
11.21 Plot Salinity Against DO

```
SalinitybyDO <-
  ggplot(OahuDataClean2, aes(x =Salinity, y =DO)) +
  geom_point(color="tomato3", alpha=1, size=0.5) +
  geom_smooth(method=lm, color="black", se=FALSE) +
  labs(title="The Effect of Salinity on DO Concentrations across Oahu", x="Salinity (ppt
print(SalinitybyDO)
```



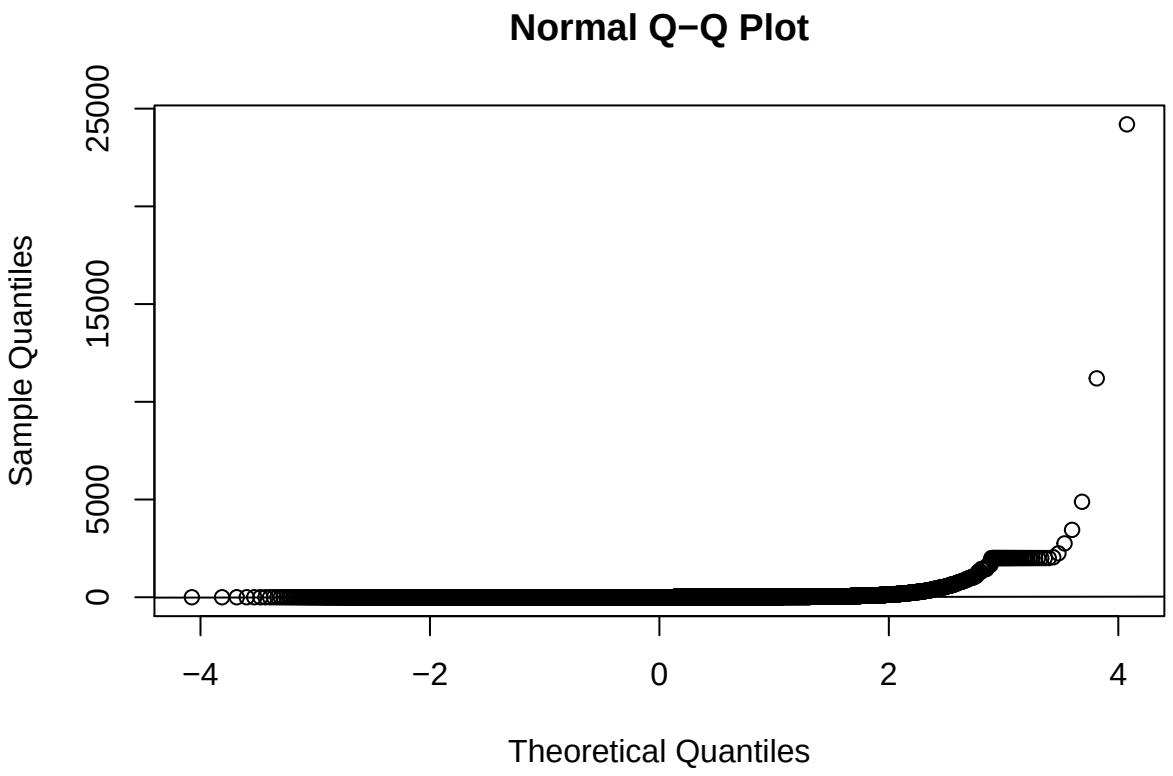
11.22 Enterococcus

```
ggplot(OahuDataClean2) +
  geom_histogram(aes(x =Enterococcus))
```



11.23 QQNorm of Enterococcus

```
qqnorm(OahuDataClean2$Enterococcus)
qqline(OahuDataClean2$Enterococcus)
```



11.24 Shapiro Test for Enterococcus

```
shapiro.test(OahuDataClean2$Enterococcus[0:5000])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  OahuDataClean2$Enterococcus[0:5000]
## W = 0.027629, p-value < 2.2e-16
```

11.25 Summary of Enterococcus

```
summary(OahuDataClean2$Enterococcus)
```

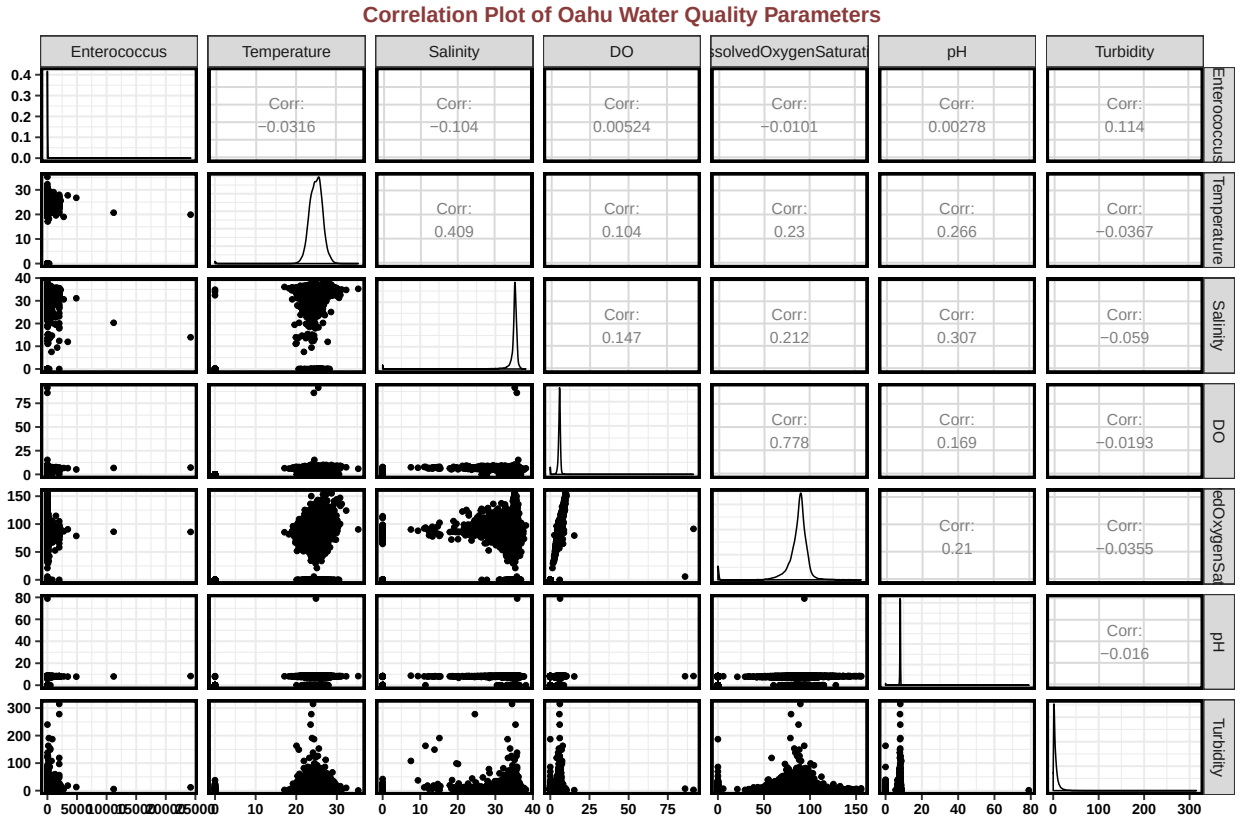
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     0.30     2.30     2.30    22.84    10.00 24196.00
```

11.25.1 Convert Date Column to a Date Object

```
OahuDataClean2$Date<-as.Date(OahuDataClean2$Date, format="%m/%d/%y")
```

12 Correlation Plot of Data

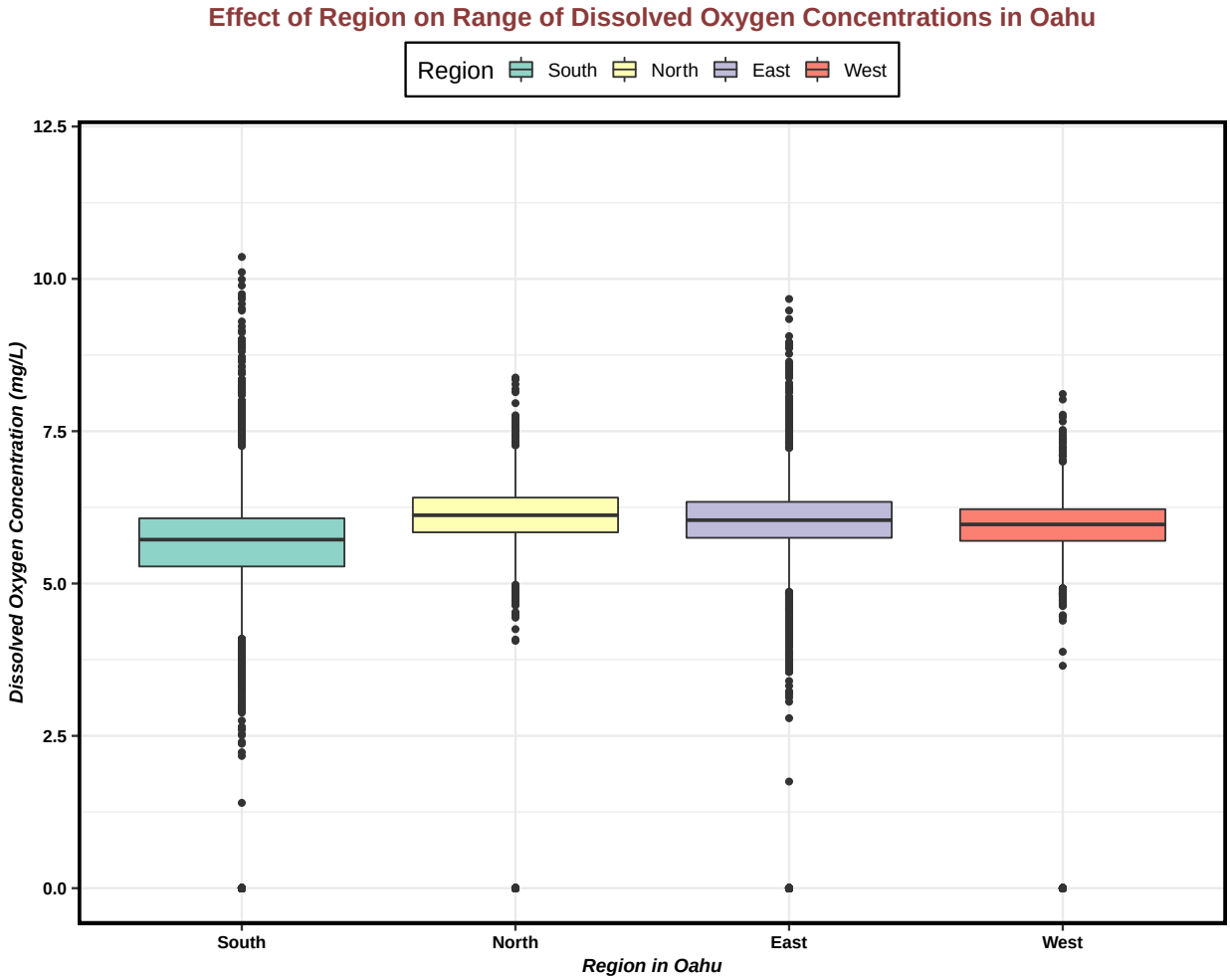
```
ggpairs(OahuDataClean, columns=c(9, 13:18), title="Correlation Plot of Oahu Water Qualit
```



13 Exploratory Boxplot showing Range of DO Concentrations by Region in Oahu

```
DOBoxplot<-ggplot(OahuDataClean) +
  geom_boxplot(aes(x=Region, y=DO, fill=Region))  +
  labs(title="Effect of Region on Range of Dissolved Oxygen Concentrations in Oahu", x="
theme(legend.title = element_text(colour="IndianRed", size=16, face="bold")) +gabytheme
scale_fill_brewer(palette="Set3") +
```

```
  scale_y_continuous(limits = c(0, 12))
print(DOBoxplot)
```



14 Statistical Analysis

14.1 Full Maximal Model

```
attach(OahuDataClean)
HawaiiModClean<-lm(DO~Enterococcus + Temperature +Salinity +pH + Turbidity + CP.Result,
summary(HawaiiModClean)
```

```
##
## Call:
## lm(formula = DO ~ Enterococcus + Temperature + Salinity + pH +
##     Turbidity + CP.Result, data = OahuDataClean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.649  -0.275   0.125   0.467  85.579
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.430e+00  1.262e-01  19.253  < 2e-16 ***
## Enterococcus  1.108e-04  4.419e-05   2.507   0.0122 *
## Temperature   1.946e-02  4.737e-03   4.108 4.01e-05 ***
## Salinity      4.203e-02  3.205e-03  13.112  < 2e-16 ***
## pH            1.709e-01  8.990e-03  19.008  < 2e-16 ***
## Turbidity    -2.084e-03  1.134e-03  -1.837   0.0662 .
## CP.Result    -1.602e-04  1.182e-03  -0.136   0.8922
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.312 on 22903 degrees of freedom
## Multiple R-squared:  0.0397, Adjusted R-squared:  0.03944
## F-statistic: 157.8 on 6 and 22903 DF,  p-value: < 2.2e-16
```

14.2 Remove Specific Heat Parameter

```
HawaiiModClean2<-update(HawaiiModClean,.~.-CP.Result)
summary(HawaiiModClean2)
```

```
##
## Call:
## lm(formula = DO ~ Enterococcus + Temperature + Salinity + pH +
##     Turbidity, data = OahuDataClean)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.648  -0.275   0.125   0.467  85.579
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.430e+00  1.261e-01  19.262  < 2e-16 ***
## Enterococcus  1.087e-04  4.145e-05   2.623  0.00873 **
## Temperature   1.946e-02  4.737e-03   4.108    4e-05 ***
## Salinity      4.204e-02  3.203e-03  13.124  < 2e-16 ***
## pH            1.709e-01  8.989e-03  19.008  < 2e-16 ***
## Turbidity    -2.116e-03  1.110e-03  -1.906  0.05667 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.312 on 22904 degrees of freedom
## Multiple R-squared:  0.03969,    Adjusted R-squared:  0.03948
## F-statistic: 189.3 on 5 and 22904 DF,  p-value: < 2.2e-16
```

14.3 Remove Turbidity Parameter

```
HawaiiModClean3<-update(HawaiiModClean2,.~.-Turbidity)
summary(HawaiiModClean3)
```

```
##
## Call:
## lm(formula = DO ~ Enterococcus + Temperature + Salinity + pH,
##     data = OahuDataClean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.639  -0.277   0.125   0.467  85.585
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.407e+00  1.256e-01  19.167  < 2e-16 ***
## Enterococcus 1.001e-04  4.121e-05   2.430   0.0151 *
## Temperature  1.960e-02  4.737e-03   4.137 3.53e-05 ***
## Salinity     4.226e-02  3.201e-03  13.202  < 2e-16 ***
## pH           1.709e-01  8.990e-03  19.006  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.312 on 22905 degrees of freedom
## Multiple R-squared:  0.03954,    Adjusted R-squared:  0.03937
## F-statistic: 235.7 on 4 and 22905 DF,  p-value: < 2.2e-16
```

14.4 AIC Test of all models

```
AIC(HawaiiModClean, HawaiiModClean2, HawaiiModClean3)
```

```
##                  df      AIC
## HawaiiModClean    8 77462.89
## HawaiiModClean2   7 77460.91
## HawaiiModClean3   6 77462.54
```

14.5 Partial F-test of all Models

```
anova(HawaiiModClean, HawaiiModClean2, HawaiiModClean3)
```

```
## Analysis of Variance Table
##
## Model 1: DO ~ Enterococcus + Temperature + Salinity + pH + Turbidity +
##     CP.Result
## Model 2: DO ~ Enterococcus + Temperature + Salinity + pH + Turbidity
## Model 3: DO ~ Enterococcus + Temperature + Salinity + pH
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1  22903 39416
## 2  22904 39416 -1   -0.0316 0.0184 0.89216
## 3  22905 39423 -1   -6.2513 3.6324 0.05668 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

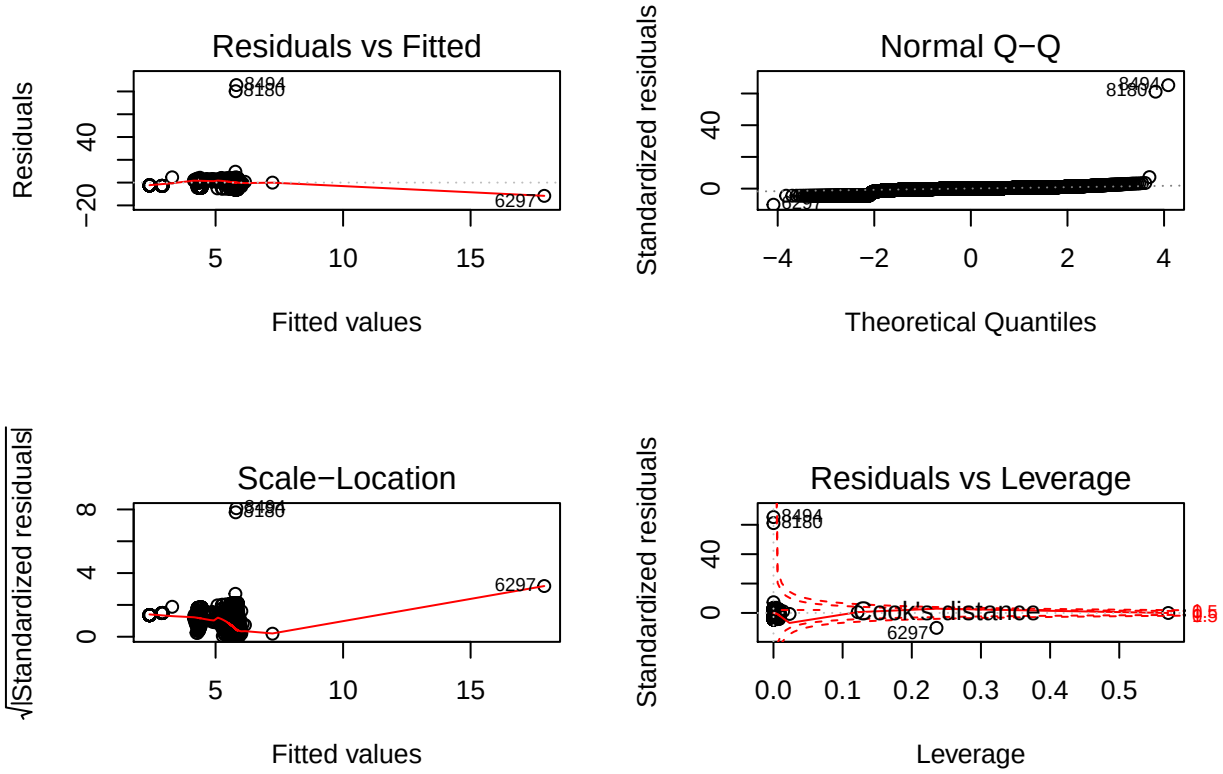
14.6 Check for Multicollinearity of Final Model

```
vif(HawaiiModClean3)
```

```
## Enterococcus  Temperature     Salinity           pH
##     1.012229     1.232857     1.279092     1.134850
```

14.6.1 Check Residuals of HawaiiModClean3

```
par(mfrow=c(2,2))
plot(HawaiiModClean3)
```



15 Research question: Is there a trend over time in DO concentrations by region in Oahu?

15.0.1 Split Dataset by Region (Use full dataset)

```
HawaiiWaterCleanOahuNorth<- filter(OahuDataClean, Region== "North")
HawaiiWaterCleanOahuSouth<- filter(OahuDataClean, Region== "South")
HawaiiWaterCleanOahuEast<- filter(OahuDataClean, Region== "East")
HawaiiWaterCleanOahuWest<- filter(OahuDataClean, Region== "West")
```

15.1 Run a Mann Kendall Test for North Oahu

```r
library(trend)
mk.test(HawaiiWaterCleanOahuNorth$DO)
```

```
##
##  Mann-Kendall trend test
##
## data:  HawaiiWaterCleanOahuNorth$DO
## z = -21.099, n = 2638, p-value < 2.2e-16
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##             S          varS           tau
## -9.531430e+05  2.040778e+09 -2.749711e-01
```

15.2 Run a Mann Kendall Test for South Oahu

```r
library(trend)
mk.test(HawaiiWaterCleanOahuSouth$DO)
```

```
##
##  Mann-Kendall trend test
##
## data:  HawaiiWaterCleanOahuSouth$DO
## z = -28.432, n = 12262, p-value < 2.2e-16
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##             S          varS           tau
## -1.286915e+07  2.048692e+11 -1.716364e-01
```

15.3 Run a Mann Kendall Test for East Oahu

```r
library(trend)
mk.test(HawaiiWaterCleanOahuEast$DO)
```

```
##
##  Mann-Kendall trend test
##
## data:  HawaiiWaterCleanOahuEast$DO
## z = -19.186, n = 4694, p-value < 2.2e-16
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##             S          varS           tau
## -2.056955e+06  1.149472e+10 -1.873351e-01
```

15.4 Run a Mann Kendall Test for West Oahu

```
library(trend)
mk.test(HawaiiWaterCleanOahuWest$DO)
```

```
##
##  Mann-Kendall trend test
##
## data:  HawaiiWaterCleanOahuWest$DO
## z = -21.833, n = 3316, p-value < 2.2e-16
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##             S          varS           tau
## -1.389943e+06  4.052891e+09 -2.538011e-01
```

For North Oahu, the z-value is 17.91, so we see a negative trend in DO concentrations over time. The p-value is $<$ 2.2e-16,so we reject the null hypothesis that the data come from a population with independent realizations and are identically distributed . For South Oahu, the z-value is -28.477, so we see a negative trend in DO concentrations over time. The p-value for South Oahu is listed as 9.14e-13, so we reject the null hypothesis that the data come from a population with independent realizations and are identically distributed. For East Oahu, the z-value is -23.022, and p-value is $<$ 2.2e-16, so we reject the null hypothesis that the data come from a population with independent realizations and are identically distributed; there is a negative trend in DO concentrations over time. For West Oahu, the z-value is -21.849, so there is also a negative trend in DO concentrations over time.

16 North Oahu

16.0.1 Arrange North Oahu Dataset by Date (ascending)

```
HawaiiWaterCleanOahuNorth<-arrange(HawaiiWaterCleanOahuNorth, Date)
```

16.1 Pettit's Test for North Oahu

```
pettitt.test(HawaiiWaterCleanOahuNorth$DO)
```

```
##
##  Pettitt's test for single change-point detection
##
## data:  HawaiiWaterCleanOahuNorth$DO
## U* = 508840, p-value < 2.2e-16
```

```
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                             1716
```

Because the p-value is <0.05, the change point is significant. Given 1st change point for North Oahu is 1332, we scroll to observation 1332 in data set, so first change point occurred in 2009-01-12.

16.2 Run a separate Mann-Kendall Test for Each Change Point

```
mk.test(HawaiiWaterCleanOahuNorth$DO[1:1331])
```

```
##
##  Mann-Kendall trend test
##
## data:  HawaiiWaterCleanOahuNorth$DO[1:1331]
## z = 10.561, n = 1331, p-value < 2.2e-16
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 1.710250e+05 2.622598e+08 1.939526e-01
```

```
mk.test(HawaiiWaterCleanOahuNorth$DO[1332:2638])
```

```
##
##  Mann-Kendall trend test
##
## data:  HawaiiWaterCleanOahuNorth$DO[1332:2638]
## z = -11.825, n = 1307, p-value < 2.2e-16
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##             S          varS           tau
## -1.863460e+05  2.483412e+08 -2.190577e-01
```

Second change point p-value<0.05, so there is a significant second change point.

17 What is the second change point?

```
pettitt.test(HawaiiWaterCleanOahuNorth$DO[1332:2638])
```

```
##
##  Pettitt's test for single change-point detection
##
```

```
## data:  HawaiiWaterCleanOahuNorth$DO[1332:2638]
## U* = 157400, p-value < 2.2e-16
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                              531
```

1332 + 942=2274, so look at 2274 row in datatable to see second change point. It happened on 2017-10-19.

18 Check for a third change point

18.0.0.1 Now split dataset into three pieces

```
mk.test(HawaiiWaterCleanOahuNorth$DO[1332:2273])
```

```
##
##  Mann-Kendall trend test
##
## data:  HawaiiWaterCleanOahuNorth$DO[1332:2273]
## z = -11.017, n = 942, p-value < 2.2e-16
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##             S          varS           tau
## -1.062570e+05  9.301715e+07 -2.405171e-01
```

```
mk.test(HawaiiWaterCleanOahuNorth$DO[2274:2638])
```

```
##
##  Mann-Kendall trend test
##
## data:  HawaiiWaterCleanOahuNorth$DO[2274:2638]
## z = -0.60241, n = 365, p-value = 0.5469
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##             S          varS           tau
## -1.404000e+03  5.424163e+06 -2.122081e-02
```

If z-score is positive, it's a positive trend in DO concentrations over time. If z-score is negative, it is a negative trend in DO concentrations over time. There is a significant trend for rows:2856:3271 because the p-value is below 0.05, so there is a third changepoint.

19 What is third change point?

```
pettitt.test(HawaiiWaterCleanOahuNorth$DO[2274:2638])
```

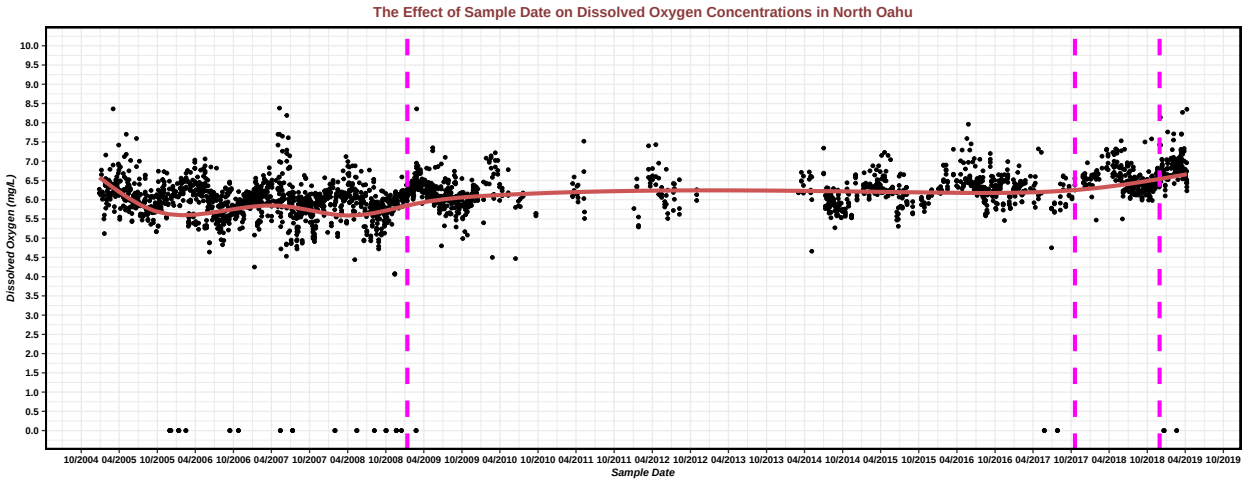
```
##
##  Pettitt's test for single change-point detection
##
## data:  HawaiiWaterCleanOahuNorth$DO[2274:2638]
## U* = 3555, p-value = 0.4223
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                             115
```

2274+215=2489–>Third change point happened on 2018-11-29

19.0.1 Change Date to be a date object

```
HawaiiWaterCleanOahuNorth$Date<-as.Date(HawaiiWaterCleanOahuNorth$Date, format = "%m/%d/
```

20 Time Series of DO Concentrations in North Oahu with Changepoints



21 South Oahu

21.0.1 Change Date to be a date object

```
HawaiiWaterCleanOahuSouth$Date<-as.Date(HawaiiWaterCleanOahuSouth$Date, format = "%m/%d/
```

21.0.2 Arrange South Oahu Dataset by Date (ascending)

```
HawaiiWaterCleanOahuSouth<-dplyr::arrange(HawaiiWaterCleanOahuSouth, Date)
```

21.1 Pettit's Test for South Oahu

```
pettitt.test(HawaiiWaterCleanOahuSouth$DO)
```

```
##
##  Pettitt's test for single change-point detection
##
## data:  HawaiiWaterCleanOahuSouth$DO
## U* = 12103000, p-value < 2.2e-16
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                            9303
```

Because the p-value is <0.05, the change point is significant. Given 1st change point for South Oahu is 9303, we scroll to observation 9303 in data set, so first change point occurred in 2012-12-12

21.2 Run a separate Mann-Kendall Test for Each Change Point

```
mk.test(HawaiiWaterCleanOahuSouth$DO[1:9302])
```

```
##
##  Mann-Kendall trend test
##
## data:  HawaiiWaterCleanOahuSouth$DO[1:9302]
## z = -2.1972, n = 9302, p-value = 0.02801
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##             S          varS           tau
```

```
## -6.571130e+05  8.944082e+10 -1.523163e-02
```

```
mk.test(HawaiiWaterCleanOahuSouth$DO[9303:12262])
```

```
##
##  Mann-Kendall trend test
##
## data:  HawaiiWaterCleanOahuSouth$DO[9303:12262]
## z = 26.793, n = 2960, p-value < 2.2e-16
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 1.438590e+06 2.882930e+09 3.293585e-01
```

p-value for [9303:12262] is significant, so run a Pettit's Test for second change point

21.3 What is second change point?

```
pettitt.test(HawaiiWaterCleanOahuSouth$DO[9303:12262])
```

```
##
##  Pettitt's test for single change-point detection
##
## data:  HawaiiWaterCleanOahuSouth$DO[9303:12262]
## U* = 1287300, p-value < 2.2e-16
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                            1922
```

9303 + 1922=11235, scroll to observation 11225 in data set, so second change point occurred in 2017-11-28

21.4 Run another Mann-Kendall to check for second change point

21.4.0.1 Now split dataset into three pieces

```
mk.test(HawaiiWaterCleanOahuSouth$DO[9303:11224])
```

```
##
##  Mann-Kendall trend test
##
## data:  HawaiiWaterCleanOahuSouth$DO[9303:11224]
## z = 3.5054, n = 1922, p-value = 0.000456
## alternative hypothesis: true S is not equal to 0
```

```
## sample estimates:
##            S         varS          tau
## 9.849200e+04 7.894562e+08 5.352165e-02
```

```
mk.test(HawaiiWaterCleanOahuSouth$DO[11225:12262])
```

```
##
##  Mann-Kendall trend test
##
## data:  HawaiiWaterCleanOahuSouth$DO[11225:12262]
## z = 4.7353, n = 1038, p-value = 2.187e-06
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 5.282400e+04 1.244360e+08 9.844092e-02
```

Because p<0.05, there is a significant change point in rows [11225:12262]

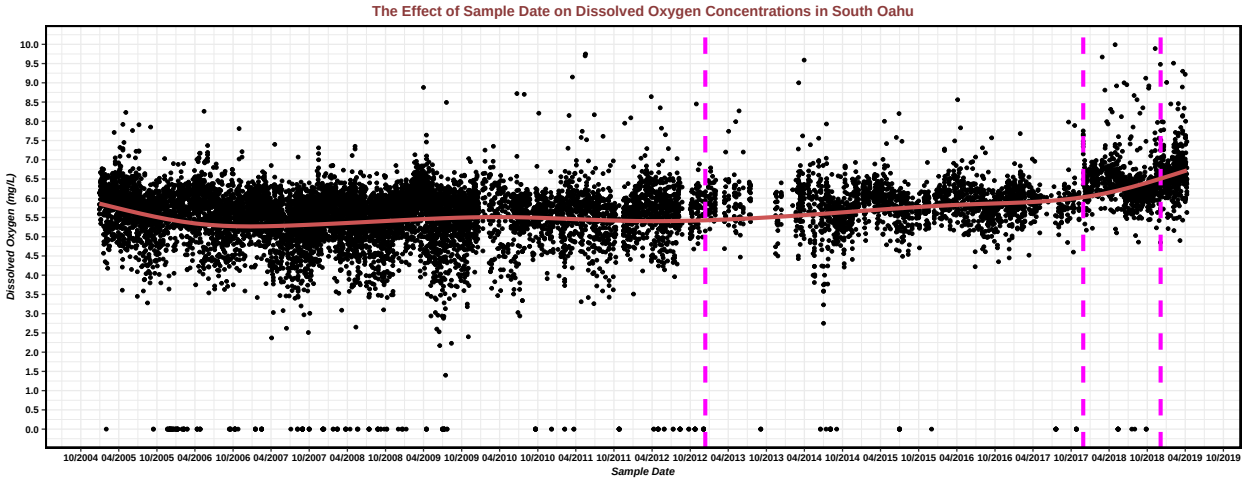
21.5 Third Change Point

```
pettitt.test(HawaiiWaterCleanOahuSouth$DO[11225:12262])
```

```
##
##  Pettitt's test for single change-point detection
##
## data:  HawaiiWaterCleanOahuSouth$DO[11225:12262]
## U* = 81125, p-value = 9.591e-16
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                             692
```

11225+ 692=11917–>third change point occurred 2018-12-04

22 Time Series of DO Concentrations in South Oahu with Changepoints



23 West Oahu

23.0.1 Change Date to be a date object

```
HawaiiWaterCleanOahuWest$Date<-as.Date(HawaiiWaterCleanOahuWest$Date, format = "%m/%d/%y
```

23.0.2 Arrange West Oahu Dataset by Date (ascending)

```
HawaiiWaterCleanOahuWest<-dplyr::arrange(HawaiiWaterCleanOahuWest, Date)
```

23.1 Pettit's Test for West Oahu

```
pettitt.test(HawaiiWaterCleanOahuWest$DO)
```

```
##
##  Pettitt's test for single change-point detection
##
## data:  HawaiiWaterCleanOahuWest$DO
## U* = 1180500, p-value < 2.2e-16
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
```

```
##                                1752
```

Because the p-value is <0.05, the change point is significant. Given 1st change point for West Oahu is 1752, we scroll to observation 1752 in data set, so first change point occurred in 2009-01-20

23.2 Run a separate Mann-Kendall Test for Each Change Point

```
mk.test(HawaiiWaterCleanOahuWest$DO[1:1751])
```

```
##
##  Mann-Kendall trend test
##
## data:  HawaiiWaterCleanOahuWest$DO[1:1751]
## z = -5.4424, n = 1751, p-value = 5.258e-08
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##             S          varS           tau
## -1.329730e+05  5.969618e+08 -8.712235e-02
```

```
mk.test(HawaiiWaterCleanOahuWest$DO[1752:3316])
```

```
##
##  Mann-Kendall trend test
##
## data:  HawaiiWaterCleanOahuWest$DO[1752:3316]
## z = 16.683, n = 1565, p-value < 2.2e-16
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 3.444320e+05 4.262618e+08 2.825216e-01
```

p-value for [1752:3316] is significant, so run a Pettit's Test for second change point

23.3 What is second change point?

```
pettitt.test(HawaiiWaterCleanOahuWest$DO[1752:3316])
```

```
##
##  Pettitt's test for single change-point detection
##
## data:  HawaiiWaterCleanOahuWest$DO[1752:3316]
## U* = 318050, p-value < 2.2e-16
## alternative hypothesis: two.sided
```

```
## sample estimates:
## probable change point at time K
##                             1118
```

1752+ 1118=2870, scroll to observation 2870 in data set, so second change point occurred in 2017-11-29

23.4 Run another Mann-Kendall to check for third change point

23.4.0.1 Now split dataset into three pieces

```
mk.test(HawaiiWaterCleanOahuWest$DO[1752:2869])
```

```
##
##  Mann-Kendall trend test
##
## data:  HawaiiWaterCleanOahuWest$DO[1752:2869]
## z = 0.52501, n = 1118, p-value = 0.5996
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 6.547000e+03 1.554565e+08 1.053231e-02
```

```
mk.test(HawaiiWaterCleanOahuWest$DO[2870:3316])
```

```
##
##  Mann-Kendall trend test
##
## data:  HawaiiWaterCleanOahuWest$DO[2870:3316]
## z = 6.2852, n = 447, p-value = 3.275e-10
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 1.983200e+04 9.955371e+06 1.998033e-01
```

Because p<0.05, there is a significant change point in rows [2870:3316]

23.5 Third Change Point

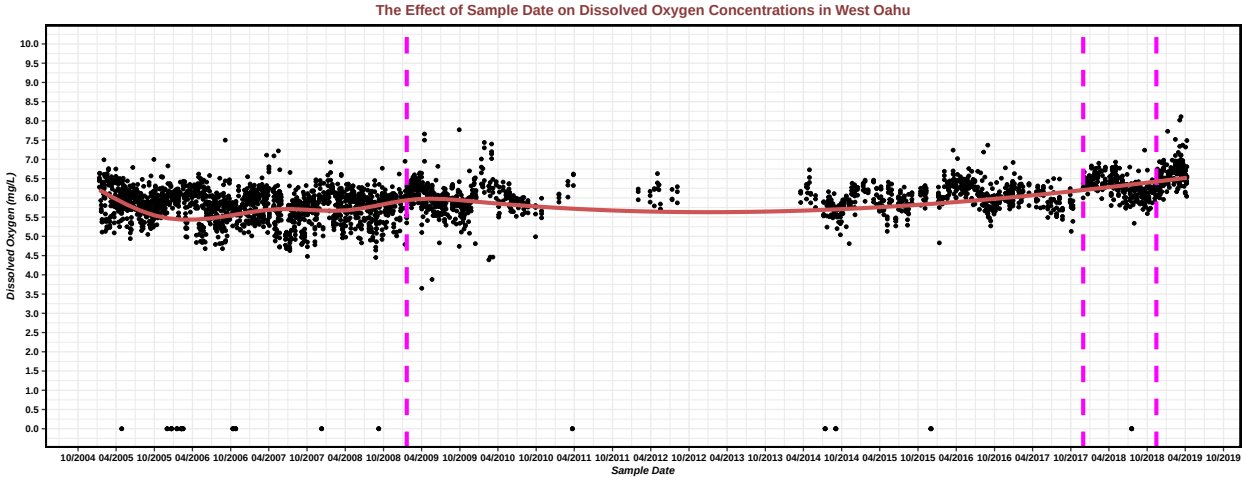
```
pettitt.test(HawaiiWaterCleanOahuWest$DO[2870:3316])
```

```
##
##  Pettitt's test for single change-point detection
##
## data:  HawaiiWaterCleanOahuWest$DO[2870:3316]
```

```
## U* = 30386, p-value < 2.2e-16
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                             273
```

Third change point is 2870+ 273=3143, which happened on 2018-11-14

24 Time Series of DO Concentrations in West Oahu with Changepoints



25 East Oahu

25.0.1 Change Date to be a date object

```
HawaiiWaterCleanOahuEast$Date<-as.Date(HawaiiWaterCleanOahuEast$Date, format = "%m/%d/%y
```

25.0.2 Arrange East Oahu Dataset by Date (ascending)

```
HawaiiWaterCleanOahuEast<-dplyr::arrange(HawaiiWaterCleanOahuEast, Date)
```

25.1 Pettit's Test for East Oahu

```
pettitt.test(HawaiiWaterCleanOahuEast$DO)
```

```
##
##  Pettitt's test for single change-point detection
##
## data:  HawaiiWaterCleanOahuEast$DO
## U* = 1848000, p-value < 2.2e-16
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                             3549
```

Because the p-value is <0.05, the change point is significant. Given 1st change point for East Oahu is 3549, we scroll to observation 3549 in data set, so first change point occurred in 2014-11-17

25.2 Run a separate Mann-Kendall Test for Each Change Point

```
mk.test(HawaiiWaterCleanOahuEast$DO[1:3548])
```

```
##
##  Mann-Kendall trend test
##
## data:  HawaiiWaterCleanOahuEast$DO[1:3548]
## z = -0.1326, n = 3548, p-value = 0.8945
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##             S          varS           tau
## -9.344000e+03  4.964278e+09 -1.490000e-03
```

```
mk.test(HawaiiWaterCleanOahuEast$DO[3549:4694])
```

```
##
##  Mann-Kendall trend test
##
## data:  HawaiiWaterCleanOahuEast$DO[3549:4694]
## z = 17.055, n = 1146, p-value < 2.2e-16
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 2.206860e+05 1.674337e+08 3.375283e-01
```

p-value for [3549:4694] is significant, so run a Pettit's Test

25.3 What is second change point?

```
pettitt.test(HawaiiWaterCleanOahuEast$DO[3549:4694])
```

```
##
##  Pettitt's test for single change-point detection
##
## data:  HawaiiWaterCleanOahuEast$DO[3549:4694]
## U* = 198770, p-value < 2.2e-16
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                              614
```

3549+ 614=4163, so look at 4163th row-> second change point occurred on 2017-11-27

26 Run another Mann-Kendall for the third change point

26.0.0.1 Now split dataset into three pieces

```
mk.test(HawaiiWaterCleanOahuEast$DO[3549:4162])
```

```
##
##  Mann-Kendall trend test
##
## data:  HawaiiWaterCleanOahuEast$DO[3549:4162]
## z = 0.56625, n = 614, p-value = 0.5712
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 2.876000e+03 2.577866e+07 1.534538e-02
```

```
mk.test(HawaiiWaterCleanOahuEast$DO[4163:4694])
```

```
##
##  Mann-Kendall trend test
##
## data:  HawaiiWaterCleanOahuEast$DO[4163:4694]
## z = 4.6476, n = 532, p-value = 3.359e-06
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 1.903600e+04 1.677464e+07 1.353027e-01
```

Because p<0.05, there is a significant change point in rows [4163:4694]

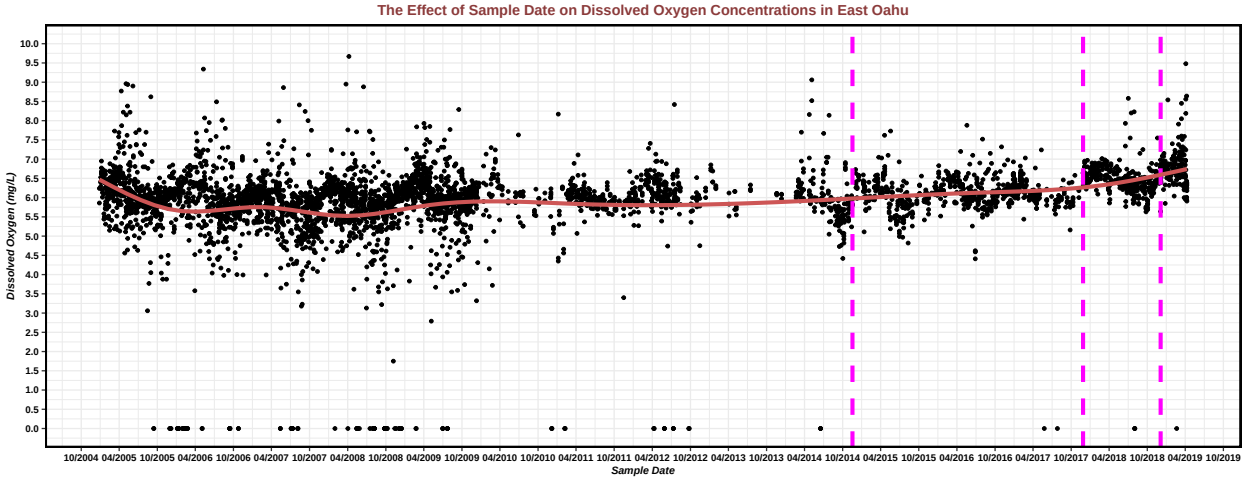
26.1 Third Change Point

```
pettitt.test(HawaiiWaterCleanOahuEast$DO[4163:4694])
```

```
##
##  Pettitt's test for single change-point detection
##
## data:  HawaiiWaterCleanOahuEast$DO[4163:4694]
## U* = 34049, p-value < 2.2e-16
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                               351
```

Third change point is 4163+ 351=4514, which happened on 2018-12-04

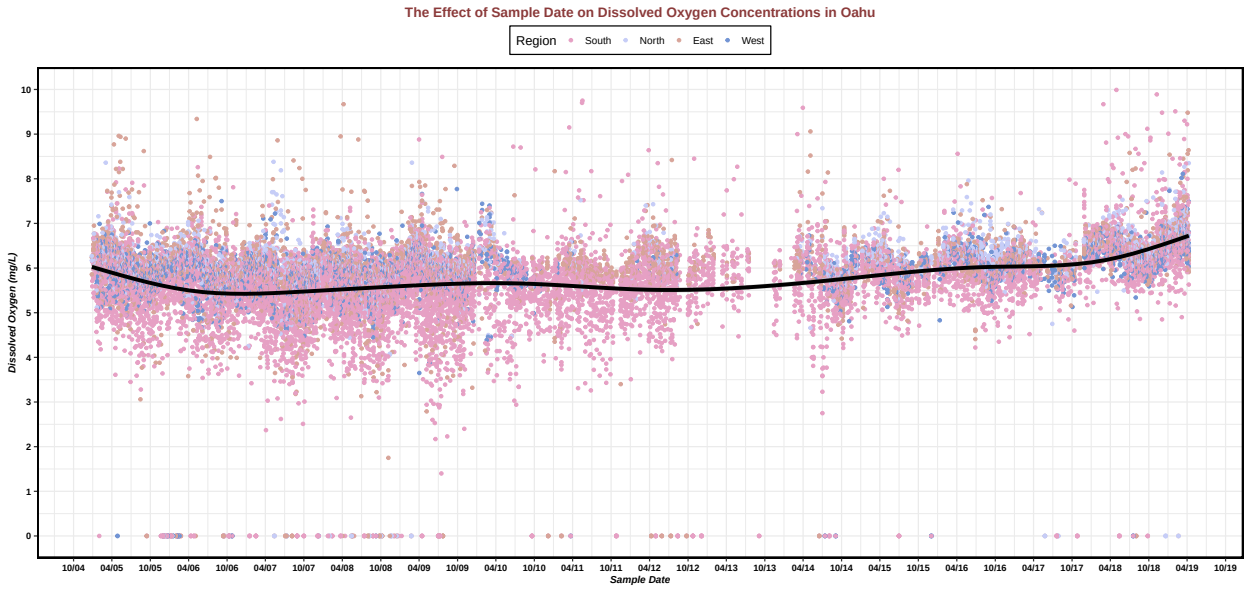
27 Time Series of DO Concentrations in East Oahu with Changepoints



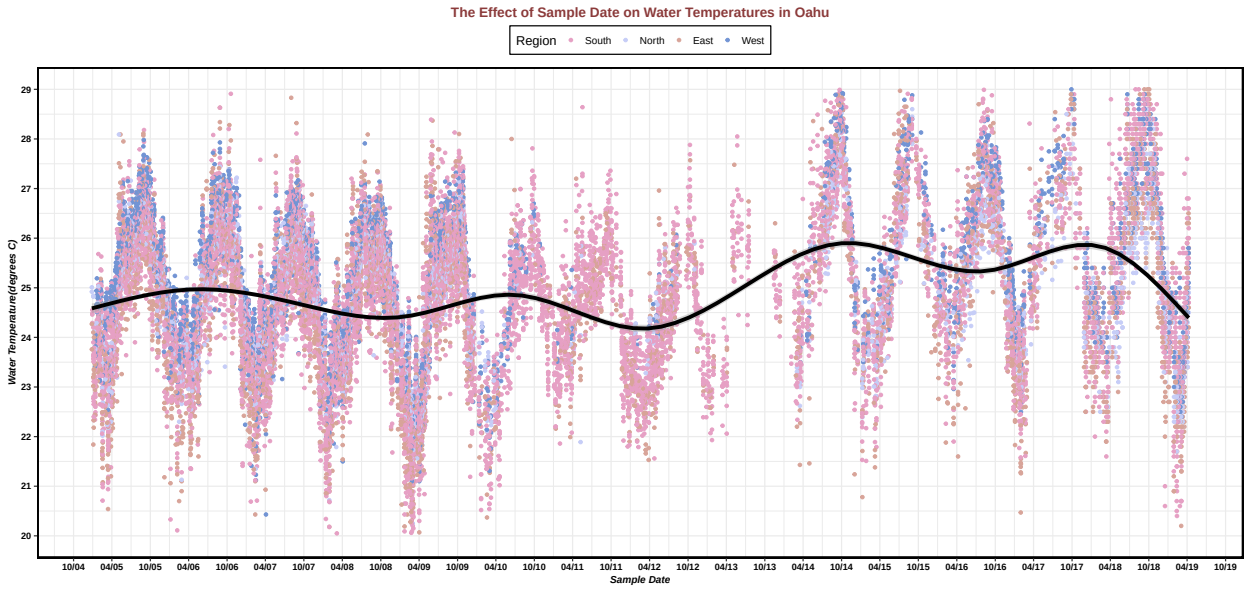
28 Effect of Sample Date on Dissolved Oxygen Concentration

28.0.1 Change Date to Date Object

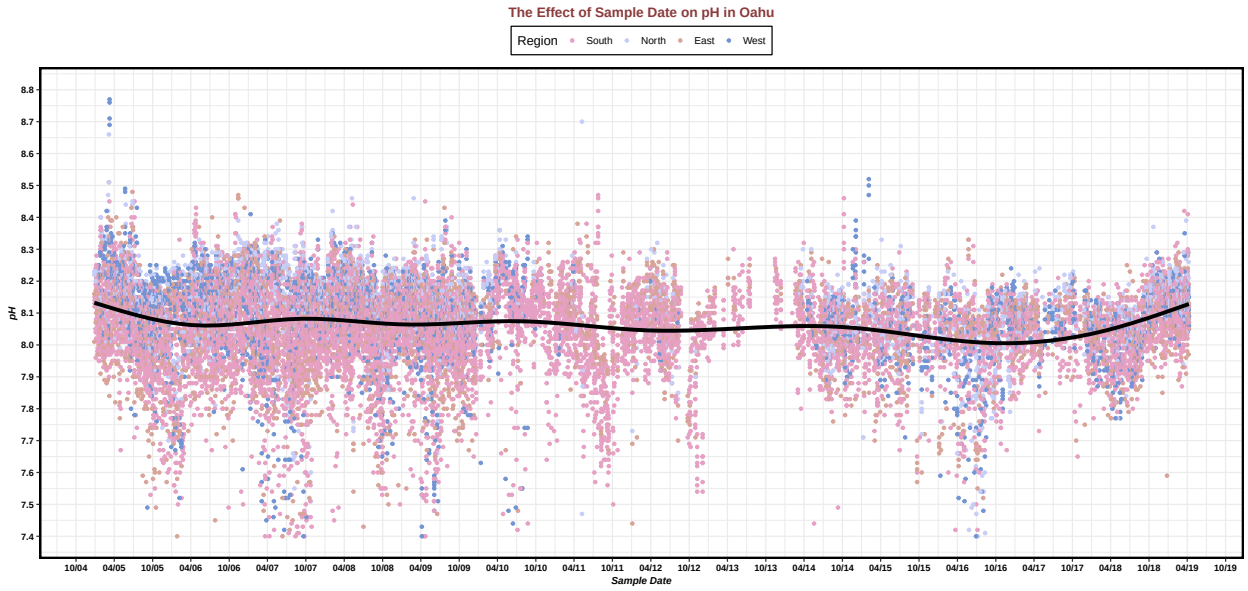
```
OahuDataClean$Date<-as.Date(OahuDataClean$Date,format = "%m/%d/%y")
```



29 Effect of Sample Date on Water Temperature in Oahu

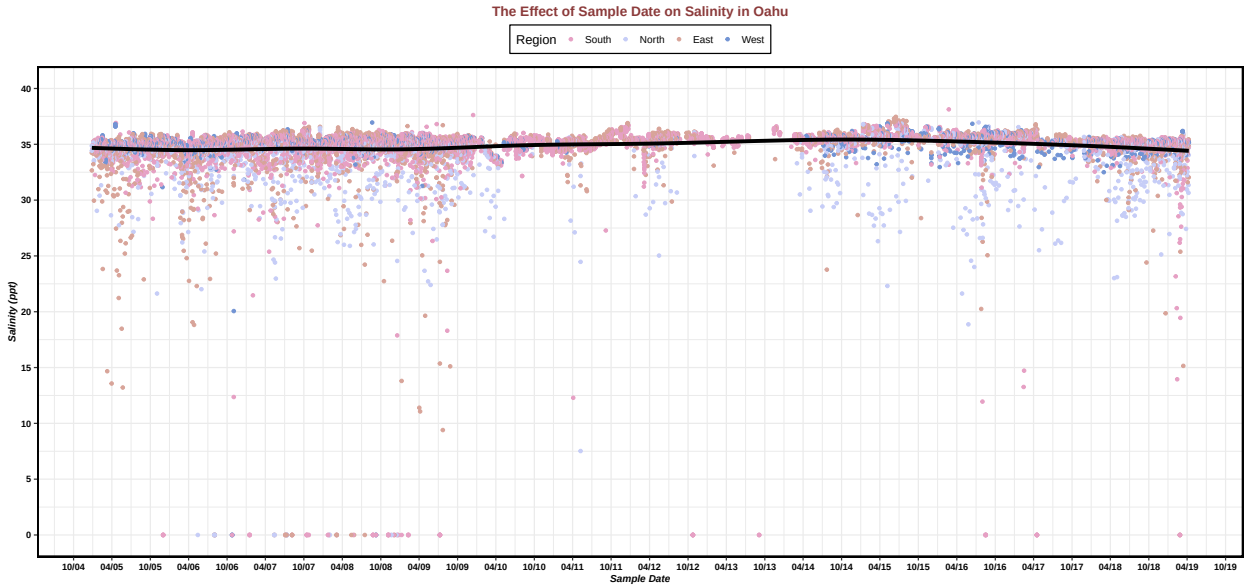


30 Effect of Sample Date on pH in Oahu

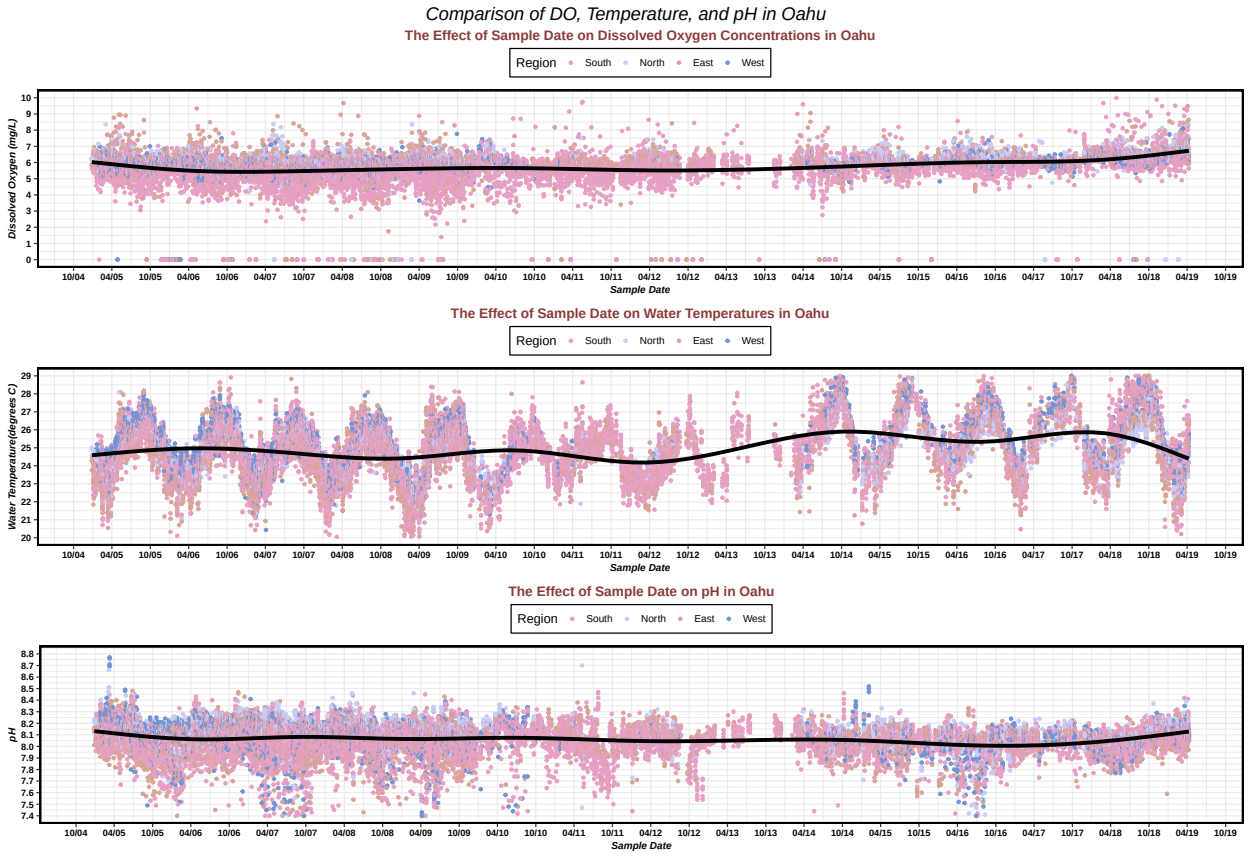


31 Effect of Sample Date on Salinity in Oahu

```
library(wesanderson)
OahuSalinityPlot<-ggplot(OahuDataClean, aes(x = Date, y =Salinity, color = Region)) +
  geom_point(alpha=1) +
scale_color_manual(values=wes_palette(name="GrandBudapest2")) +
  ##geom_vline(xintercept=as.Date("2004-08-16"),color="253494", origin= "1970-01-01", lt
 ## geom_vline(xintercept=as.Date("2005-08-17"), color="253494", origin= "1970-01-01", l

  geom_smooth(aes(x = Date, y =Salinity, span=0.1), color="black", linetype=1, size=2) +
  labs(title="The Effect of Sample Date on Salinity in Oahu",
       x="Sample Date",
       y="Salinity (ppt)") +

 scale_x_date(labels = date_format("%m/%y"), breaks = date_breaks("6 month")) +
scale_y_continuous(limits=c(0,40), breaks=seq(0, 40, by = 5))


OahuSalinityPlot
```



32 Comparison of pH, DO and Temperature over Time in Oahu



33 Spatial Analysis

33.1 Compute Mean DO, Mean Temp, and Mean Salinity by Oahu Sample Location

```
OahuDataCleanAvg<-OahuDataClean%>%
   group_by(LocationName, LocationIdentifier, Region, LatDecDeg, LongDecDeg) %>%
  summarize(meanDO= mean(DO),
            meanTemp = mean(Temperature),
            meanTurbidity=mean(Turbidity))
```

34 Summary and Conclusions

Using exploratory plots, I found that the data is not normally distributed, like most environmental datasets. As seen by my exploratory plots, some of my variables were right skewed with some extreme values and outliers. However, assumptions of multiple linear regression are independence of error terms, homoscedasticity(constant variance) of errors, normality of the error distribution, explanatory variables are fixed, and no perfect multicollinearity. I decided to proceed with a multiple linear regression, as my residual errors looked fine. The maximal model is HawaiiModClean with all relevant independent variables included. Using my GGPairs Bivariate scatterplots for all variable combinations allowed me to look at the correlation coefficients between each variable combination, to ensure there is no multicollinearity between the independent variables, and the independent and dependent variables are linearly correlated. Enterococcus and DO have almost no correlation: 0.005. Temperature and DO have a low positive correlation: 0.104. Salinity and DO have a low positive correlation: 0.147. pH and DO have a low positive correlation: 0.169. Turbidity and DO have almost no correlation: -0.019. Percent Saturation DO and DO are multicollineated because they are related variables, so I did not include Percent Saturation of DO in my analysis.

I used the stepwise reduction methodto remove the least significant independent variables from the maximal model until only the independent variables with a significant effect on dissolved oxygen concentrations were left, leaving the most parsimonious model HawaiiModClean3. An AIC test was done comparing our final reduced model to our fuller models to determine a goodness of fit score. Our final reduced model had an AIC score 77790.15. While this wasn't the lowest model score, it was within three points of the next score, and Turbidity was not statistically significant with a p-value >0.05, so I removed it. My multiple linear regression found that the explanatory parameters of temperature, pH, salinity, and Enterococci concentrations were significant predictors of dissolved oxygen concentrations in Oahu. A VIF test found that none of these parameters suffered from multicollinearity. The following are their respective statistics: Temperature (p=4.34e-05, t=4.09), Enterococcus (t=2.42, p=0.015), Salinity (t=13.096, p<2e-16), and pH(t=19.03, p<2e-16). Based on the model coefficients, a one degree Celsius increase in temperature will result in an increase

in dissolved oxygen concentrations by 0.02 mg/L. A one ppt increase in Salinity will result in an increase in dissolved oxygen concentrations by 0.042 mg/L. A one unit increase in pH will result in an increase in dissolved oxygen concentrations by 0.17 mg/L. A one unit increase in Enterococcus will result in an increase in dissolved oyxgen concentrations by 9.99e-5 mg/L. (Linear Regression; $p<0.05$, df=15,449, Rsqaured=0.04). The linear equation for my regression is Dissolved Oxygen=2.42 + 0.02(Temperature)+0.042(Salinity) + 0.17(pH) +9.99e-5(Enterococcus) + E. If I had more time, I would have researched interactions between these parameters to see which interactions would have been the most beneficial to include in my maximal model. My model only explains 4% of the variability of the data around its mean, and perhaps certain interactions would account for more of the variability.

If I had more available parameters in my data such as phosphorus and nitrate concentrations, they might explain more of the variability in my dataset that my multiple linear regression failed to explain. Phoshorus and nitrate often end up in water bodies as runoff from agriculture/anthropogenic activity and deplete the dissolved oxygen necessary for aerobic activity in water bodies. Because Honolulu (the capital of Hawaii) is on the South coast of Oahu, I surmise that DO concentrations would be lower on the south coast if I was able to include phosphorus and nitrate measurements in my data. In addition, if I had data on marine organisms (that use up oxygen) or ocean circulation/stratification, that would also explain more of the variability.

My second research question concerned whether dissolved oxygen concentrations varied spatially across the North, South, East, and West coasts of Oahu. An exploratory boxplot showed that the median of DO concentrations on the South Coast was slightly lower than the median DO Concentrations of the North, West, and East coasts. Honolulu is located on the South coast and several samples were taken by the capital city, but this is not enough evidence to draw a conclusion.