Untitled

Brianne Bell

2022-11-05

```
library(gridExtra)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(ggplot2)

library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✓ tibble  3.1.7     ✓ dplyr   1.0.9
## ✓ tidyr   1.2.0     ✓ stringr 1.4.0
## ✓ readr   2.1.2     ✓ forcats 0.5.1
## ✓ purrr   0.3.4
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✕ dplyr::combine() masks gridExtra::combine()
## ✕ dplyr::filter()  masks stats::filter()
## ✕ dplyr::lag()     masks stats::lag()
```

Ocean Quality Dataset:

https://data.sandiego.gov/datasets/monitoring-ocean-water-quality/ (https://data.sandiego.gov/datasets/monitoring-ocean-water-quality/)

```
ocean20 <- read.csv ("C:/Users/breel.B-E-BELL/OneDrive/Documents/USD_MastersAppliedDataScience/A
DS-506 Applied Time Series/Data/water_quality_2020_2021_datasd.csv")
head(ocean20, 5)
```

```
##       sample station depth_m date_sample        time project parameter
## 1 2001018683     S11      NA  2020-01-01 9:18:00 PST    SBOO    ENTERO
## 2 2001018683     S11      NA  2020-01-01 9:18:00 PST    SBOO     TOTAL
## 3 2001018683     S11      NA  2020-01-01 9:18:00 PST    SBOO     FECAL
## 4 2001018680      S4      NA  2020-01-01                SBOO     TOTAL
## 5 2001018680      S4      NA  2020-01-01                SBOO     FECAL
##   qualifier value      units
## 1         e   220 CFU/100 mL
## 2        NR    NA CFU/100 mL
## 3        NR    NA CFU/100 mL
## 4        NS    NA CFU/100 mL
## 5        NS    NA CFU/100 mL
```

```
str(ocean20)
```

```
## 'data.frame':    70163 obs. of  10 variables:
##  $ sample     : int  2001018683 2001018683 2001018683 2001018680 2001018680 2001018682 200101
8682 2001018682 2001029022 2001029022 ...
##  $ station    : chr  "S11" "S11" "S11" "S4" ...
##  $ depth_m    : int  NA NA NA NA NA NA NA NA 1 1 ...
##  $ date_sample: chr  "2020-01-01" "2020-01-01" "2020-01-01" "2020-01-01" ...
##  $ time       : chr  "9:18:00 PST" "9:18:00 PST" "9:18:00 PST" "" ...
##  $ project    : chr  "SBOO" "SBOO" "SBOO" "SBOO" ...
##  $ parameter  : chr  "ENTERO" "TOTAL" "FECAL" "TOTAL" ...
##  $ qualifier  : chr  "e" "NR" "NR" "NS" ...
##  $ value      : num  220 NA NA NA NA ...
##  $ units      : chr  "CFU/100 mL" "CFU/100 mL" "CFU/100 mL" "CFU/100 mL" ...
```

- sample
    - unique sample ID
- station
    - unique location ID
- depth_m
    - depth in meters of sample
- date_sample
    - date sample taken
- time
    - time sample taken
- project
    - Outfall region where sample was collected
        - PLOO (PL): Point Loma Ocean Outfall;
        - SBOO: South Bay Ocean Outfall
- parameter
    - factor being recorded
        - fluorometry
        - *DENSITY (sigma-t)*
        - *DO (dissolved oxygen) (mg/L)*
        - *ENTERO (CFU/100mL)*
        - *FECAL (CFU/100mL)*

        - OG (?)
        - *PH (pH)*
        - *SALINITY (ppt)*
        - SUSO
        - *TEMP (C)*
        - TOTAL ?
        - pct_light
- qualifier
    - qualifier for value
        - <, >, e, LA, ND, NS
- value
    - actual value of measurement
- units
    - units for each factor
        - %;
        - C; (for temperature)
        - CFU/100 ml; (for entero, fecal)
        - mg/L; (for DO)
        - pH; (for pH)
        - ppt; (for salinity)
        - sigma-t; (for density)
        - ug/L (for chlorophyl)

```
sapply(ocean20, function(x) sum(is.na(x)))
```

```
##      sample     station     depth_m date_sample        time     project 
##           0           0        6833           0           0           0 
##   parameter   qualifier       value       units 
##           0           0          85           0
```

```
  # most missing values (6833 of 70163) in the depth feature
  # then only missing values in value column, unsure at the moment which measurements are missing
```

```
Entero_CFUper100mL <- ifelse(ocean20$parameter=='ENTERO', ocean20$value, 0)
Fecal_CFUper100mL <- ifelse(ocean20$parameter=='FECAL', ocean20$value, 0)
Temp_C <- ifelse(ocean20$parameter=='TEMP', ocean20$value, 0)
DO_mgperL <- ifelse(ocean20$parameter=='DO', ocean20$value, 0)
pH <- ifelse(ocean20$parameter=='PH', ocean20$value, 0)
Chlorophyll_ugperL <- ifelse(ocean20$parameter=='CHLOROPHYLL', ocean20$value, 0)
XMS_pct <- ifelse(ocean20$parameter=='XMS', ocean20$value, 0)
Salinity_ppt <- ifelse(ocean20$parameter=='SALINITY', ocean20$value, 0)
Density_sigmat <- ifelse(ocean20$parameter=='DENSITY', ocean20$value, 0)

ocean_df <- data.frame(sample=ocean20$sample,
                       station=ocean20$station,
                       depth_m=ocean20$depth_m,
                       date_sample=ocean20$date_sample,
                       time=ocean20$time,
                       project=ocean20$project,
                       Entero_CFUper100mL=Entero_CFUper100mL,
                       Fecal_CFUper100mL=Fecal_CFUper100mL,
                       Temp_C=Temp_C,
                       DO_mgperL=DO_mgperL,
                       pH=pH,
                       Chlorophyll_ugperL=Chlorophyll_ugperL,
                       XMS_pct=XMS_pct,
                       Salinity_ppt=Salinity_ppt,
                       Density_sigmat=Density_sigmat)
head(ocean_df,25)
```

```
##        sample station depth_m date_sample        time project
## 1  2001018683     S11      NA  2020-01-01 9:18:00 PST    SBOO
## 2  2001018683     S11      NA  2020-01-01 9:18:00 PST    SBOO
## 3  2001018683     S11      NA  2020-01-01 9:18:00 PST    SBOO
## 4  2001018680      S4      NA  2020-01-01                SBOO
## 5  2001018680      S4      NA  2020-01-01                SBOO
## 6  2001018682      S6      NA  2020-01-01 9:31:00 PST    SBOO
## 7  2001018682      S6      NA  2020-01-01 9:31:00 PST    SBOO
## 8  2001018682      S6      NA  2020-01-01 9:31:00 PST    SBOO
## 9  2001029022      A1       1  2020-01-02 7:44:00 PST    PLOO
## 10 2001029022      A1       1  2020-01-02 7:44:00 PST    PLOO
## 11 2001029022      A1       1  2020-01-02 7:44:00 PST    PLOO
## 12 2001029022      A1       1  2020-01-02 7:44:00 PST    PLOO
## 13 2001029022      A1       1  2020-01-02 7:44:00 PST    PLOO
## 14 2001029022      A1       1  2020-01-02 7:44:00 PST    PLOO
## 15 2001029022      A1       1  2020-01-02 7:44:00 PST    PLOO
## 16 2001028460      A1       1  2020-01-02 7:44:00 PST    PLOO
## 17 2001028460      A1       1  2020-01-02 7:44:00 PST    PLOO
## 18 2001028460      A1       1  2020-01-02 7:44:00 PST    PLOO
## 19 2001029023      A1      12  2020-01-02 7:44:00 PST    PLOO
## 20 2001029023      A1      12  2020-01-02 7:44:00 PST    PLOO
## 21 2001029023      A1      12  2020-01-02 7:44:00 PST    PLOO
## 22 2001029023      A1      12  2020-01-02 7:44:00 PST    PLOO
## 23 2001029023      A1      12  2020-01-02 7:44:00 PST    PLOO
## 24 2001029023      A1      12  2020-01-02 7:44:00 PST    PLOO
## 25 2001029023      A1      12  2020-01-02 7:44:00 PST    PLOO
##    Entero_CFUper100mL Fecal_CFUper100mL Temp_C DO_mgperL   pH
## 1                 220                 0   0.00      0.00 0.00
## 2                   0                 0   0.00      0.00 0.00
## 3                   0                NA   0.00      0.00 0.00
## 4                   0                 0   0.00      0.00 0.00
## 5                   0                NA   0.00      0.00 0.00
## 6                   0                NA   0.00      0.00 0.00
## 7                 200                 0   0.00      0.00 0.00
## 8                   0                 0   0.00      0.00 0.00
## 9                   0                 0  15.01      0.00 0.00
## 10                  0                 0   0.00      7.69 0.00
## 11                  0                 0   0.00      0.00 8.10
## 12                  0                 0   0.00      0.00 0.00
## 13                  0                 0   0.00      0.00 0.00
## 14                  0                 0   0.00      0.00 0.00
## 15                  0                 0   0.00      0.00 0.00
## 16                  0                 2   0.00      0.00 0.00
## 17                  0                 0   0.00      0.00 0.00
## 18                  2                 0   0.00      0.00 0.00
## 19                  0                 0   0.00      0.00 0.00
## 20                  0                 0   0.00      0.00 0.00
## 21                  0                 0   0.00      7.61 0.00
## 22                  0                 0   0.00      0.00 0.00
## 23                  0                 0  15.39      0.00 0.00
## 24                  0                 0   0.00      0.00 0.00
## 25                  0                 0   0.00      0.00 8.11
##    Chlorophyll_ugperL XMS_pct Salinity_ppt Density_sigmat
## 1                0.00    0.00        0.000          0.000
## 2                0.00    0.00        0.000          0.000
## 3                0.00    0.00        0.000          0.000
## 4                0.00    0.00        0.000          0.000
## 5                0.00    0.00        0.000          0.000
## 6                0.00    0.00        0.000          0.000
## 7                0.00    0.00        0.000          0.000
## 8                0.00    0.00        0.000          0.000
## 9                0.00    0.00        0.000          0.000
## 10               0.00    0.00        0.000          0.000
## 11               0.00    0.00        0.000          0.000
## 12               1.70    0.00        0.000          0.000
## 13               0.00   77.11        0.000          0.000
## 14               0.00    0.00       33.283          0.000
## 15               0.00    0.00        0.000         24.647
## 16               0.00    0.00        0.000          0.000
## 17               0.00    0.00        0.000          0.000
## 18               0.00    0.00        0.000          0.000
## 19               0.00    0.00       33.409          0.000
## 20               0.00    0.00        0.000         24.659
## 21               0.00    0.00        0.000          0.000
## 22               1.62    0.00        0.000          0.000
## 23               0.00    0.00        0.000          0.000
## 24               0.00   81.66        0.000          0.000
## 25               0.00    0.00        0.000          0.000
```

```r
# histograms
ocplots <- subset(ocean_df, select= c('Entero_CFUper100mL', 'Fecal_CFUper100mL', 'Temp_C',
                       'DO_mgperL', 'pH', 'Chlorophyll_ugperL', 'XMS_pct',
                       'Salinity_ppt', 'Density_sigmat'))

library(Hmisc)
```

```
## Warning: package 'Hmisc' was built under R version 4.2.2
```

```
## Loading required package: lattice
```

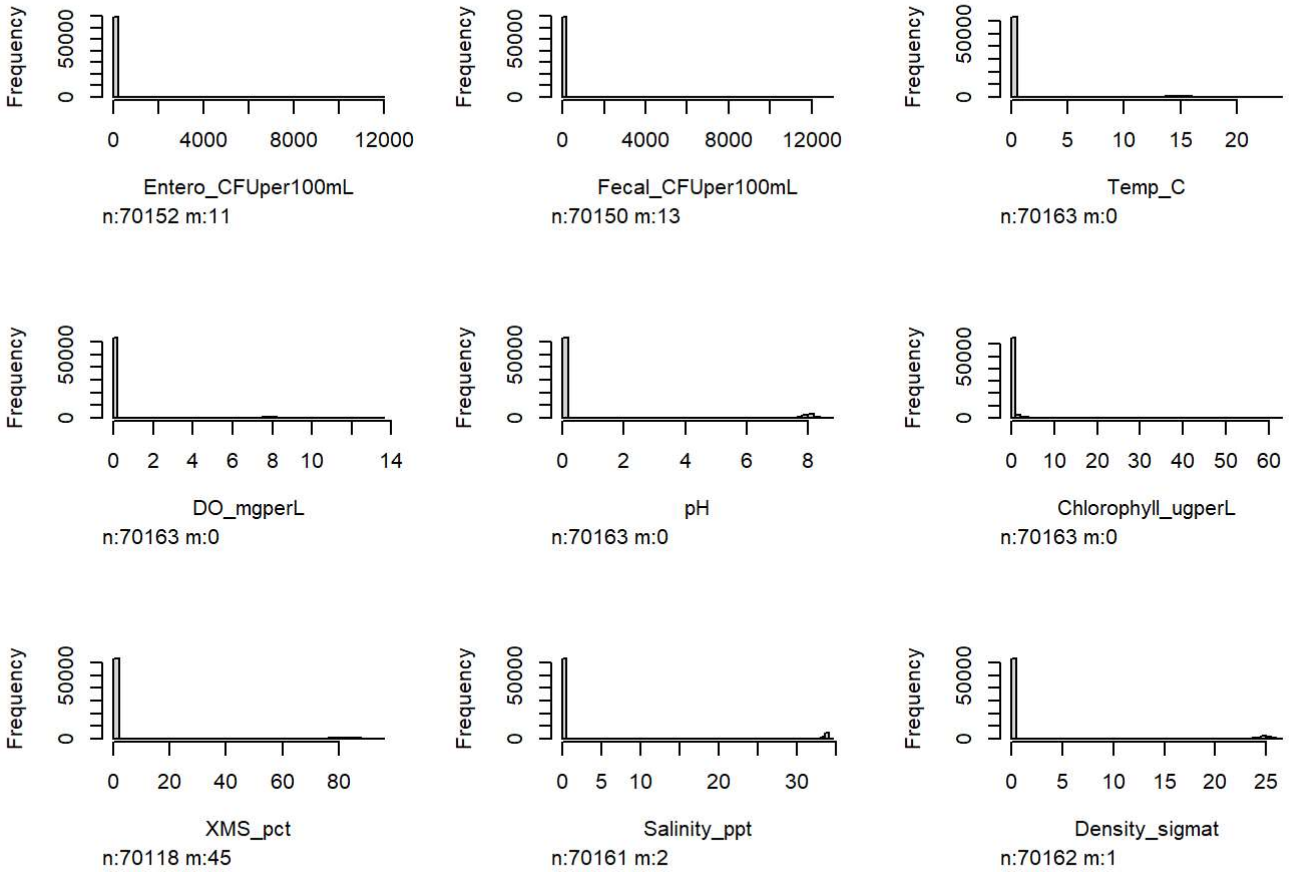
```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
##
## Attaching package: 'Hmisc'
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
hist.data.frame(ocplots)
```



```
# histogram(ocplots$Entero_CFUper100mL, main="Entero in CFU/100mL")
# histogram(ocplots$Fecal_CFUper100mL, main="Fecal in CFU/100mL")
# histogram(ocplots$Temp_C, main="Temperature (C)")
# histogram(ocplots$DO_mgperL, main="Dissolved Oxygen in mg/L")
# histogram(ocplots$pH, main="pH")
# histogram(ocplots$Chlorophyll_ugperL, main="Chlorphyll in ug/L")
# histogram(ocplots$Salinity_ppt, main="Salinity in ppt (parts per trillion")
# histogram(ocplots$Density_sigmat, main="Density in sigma-t")
# histogram(ocplots$XMS_pct, main="XMS (I don't know what it is) in percent")
```

```
summary(ocean_df)
```

```
##      sample            station             depth_m      date_sample
##  Min.   :2.001e+09   Length:70163       Min.   : 1.00   Length:70163
##  1st Qu.:2.006e+09   Class :character   1st Qu.: 2.00   Class :character
##  Median :2.012e+09   Mode  :character   Median : 9.00   Mode  :character
##  Mean   :2.056e+09                      Mean   :12.96
##  3rd Qu.:2.106e+09                      3rd Qu.:18.00
##  Max.   :2.112e+09                      Max.   :98.00
##                                         NA's   :6833
##      time             project          Entero_CFUper100mL Fecal_CFUper100mL
##  Length:70163       Length:70163       Min.   :    0.00   Min.   :    0.00
##  Class :character   Class :character   1st Qu.:    0.00   1st Qu.:    0.00
##  Mode  :character   Mode  :character   Median :    0.00   Median :    0.00
##                                        Mean   :   39.25   Mean   :   49.03
##                                        3rd Qu.:    0.00   3rd Qu.:    0.00
##                                        Max.   :12000.00   Max.   :13000.00
##                                        NA's   :11         NA's   :13
##      Temp_C         DO_mgperL             pH         Chlorophyll_ugperL
##  Min.   : 0.000   Min.   : 0.0000   Min.   :0.0000   Min.   : 0.0000
##  1st Qu.: 0.000   1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.: 0.0000
##  Median : 0.000   Median : 0.0000   Median :0.0000   Median : 0.0000
##  Mean   : 1.424   Mean   : 0.6776   Mean   :0.7603   Mean   : 0.2358
##  3rd Qu.: 0.000   3rd Qu.: 0.0000   3rd Qu.:0.0000   3rd Qu.: 0.0000
##  Max.   :23.930   Max.   :13.6000   Max.   :8.7100   Max.   :62.3100
##
##     XMS_pct        Salinity_ppt   Density_sigmat
##  Min.   : 0.000   Min.   : 0.00   Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.: 0.00   1st Qu.: 0.000
##  Median : 0.000   Median : 0.00   Median : 0.000
##  Mean   : 7.308   Mean   : 3.18   Mean   : 2.352
##  3rd Qu.: 0.000   3rd Qu.: 0.00   3rd Qu.: 0.000
##  Max.   :94.190   Max.   :34.21   Max.   :26.301
##  NA's   :45       NA's   :2       NA's   :1
```

Ocean Water:

Title:

Rising temperature and falling acidity (assuming this wtih reports of reefs being hurt by more acidic waters)

Motivation:

Climate change and warming oceans

Problem Statement:

Utilize ocean water measurements (salinity, temperature, density, chlorophyll, dissolved oxygen, and pH) to see impact over time. Ideally, predict temperature and acidity (pH) in set time period in the future, based on the measurements from 2020-2021.

Notable Findings:

We need a better way to split the “parameter” column into separate columns for each measurement so that we don’t have a ton of “zero” values as place holders.

Drinking Water Chemicals Dataset:

https://data.sandiego.gov/datasets/monitoring-of-select-chemical-parameters-in-drinking-water/ (https://data.sandiego.gov/datasets/monitoring-of-select-chemical-parameters-in-drinking-water/)

```
chem <- read.csv ("C:/Users/breel.B-E-BELL/OneDrive/Documents/USD_MastersAppliedDataScience/ADS-
506 Applied Time Series/Data/analyte_tests_drinking_water_datasd.csv")
head(chem, 5)
```

```
##   date_sample sample_source sample_id  analyte value_qualifier analyte_value
## 1  2022-01-01       55A SYS  W1470689 FLUORIDE                         0.469
## 2  2022-01-02       174 SYS  W1470694 FLUORIDE                         0.438
## 3  2022-01-03       313 SYS  W1471820 FLUORIDE                         0.478
## 4  2022-01-03       50A SYS  W1471858    COLOR              ND            NA
## 5  2022-01-03       50A SYS  W1471858      TON                         1.000
##   value_units                         source_description
## 1        MG/L                         5183 Arvinels Ave.
## 2        MG/L 3250 Camino Del Rio North; Sample Stanchion
## 3        MG/L                        11602 Calle Paracho
## 4       COLOR                         2693 Melbourne Dr.
## 5        ODOR                         2693 Melbourne Dr.
```

```
str(chem)
```

```
## 'data.frame':    20323 obs. of  8 variables:
##  $ date_sample       : chr  "2022-01-01" "2022-01-02" "2022-01-03" "2022-01-03" ...
##  $ sample_source     : chr  "55A SYS" "174 SYS" "313 SYS" "50A SYS" ...
##  $ sample_id         : chr  "W1470689" "W1470694" "W1471820" "W1471858" ...
##  $ analyte           : chr  "FLUORIDE" "FLUORIDE" "FLUORIDE" "COLOR" ...
##  $ value_qualifier   : chr  "" "" "" "ND" ...
##  $ analyte_value     : num  0.469 0.438 0.478 NA 1 0.1 NA 2 0.05 NA ...
##  $ value_units       : chr  "MG/L" "MG/L" "MG/L" "COLOR" ...
##  $ source_description: chr  "5183 Arvinels Ave." "3250 Camino Del Rio North; Sample Stanchio
n" "11602 Calle Paracho" "2693 Melbourne Dr." ...
```

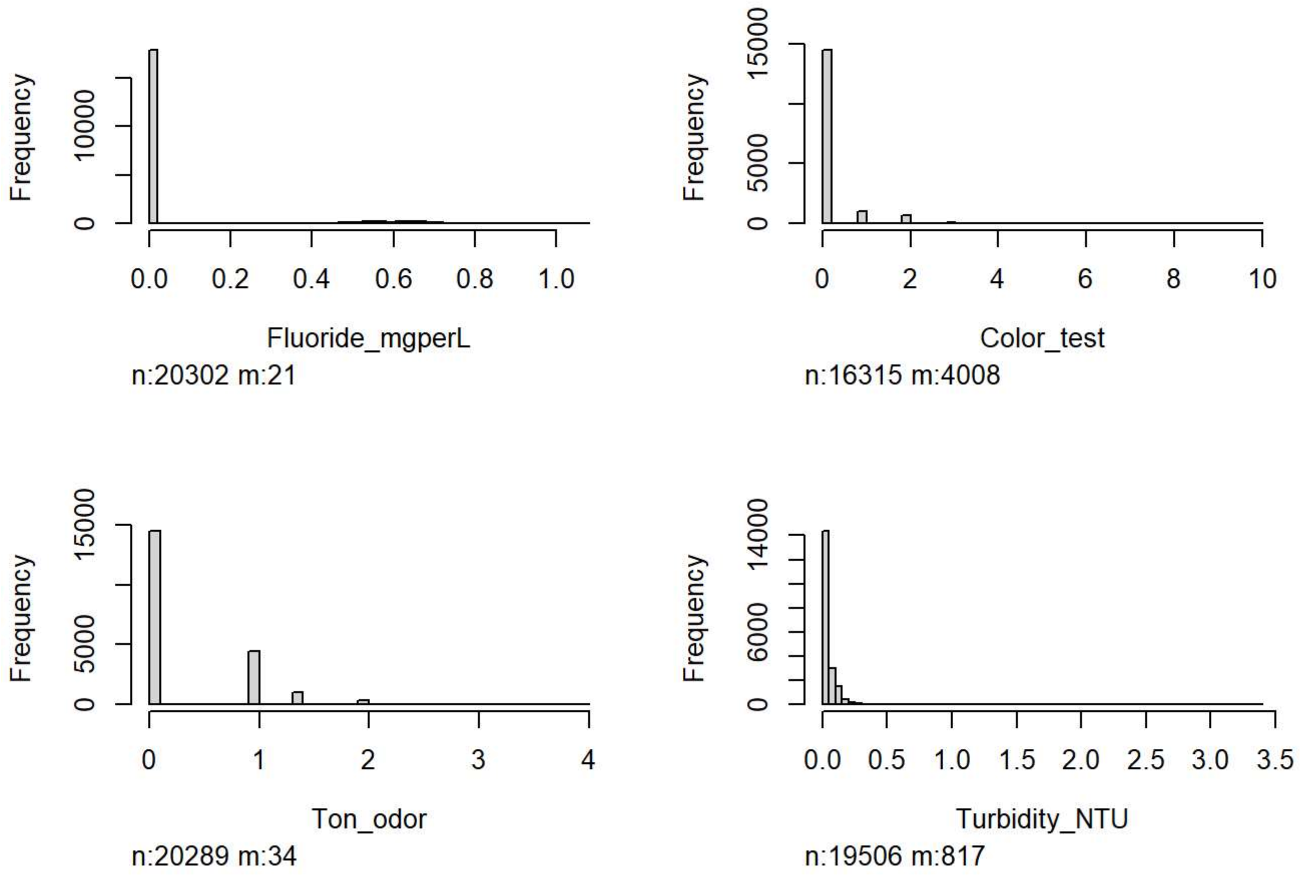
- date_sample
    - date and time sample taken
- sample_source
    - location where sample taken -sample_id
    - unique sample identifier
- analyte
    - analyte (chemical) measured (categorical, need to split into dummy columns)
        - FLUORIDE (mg/L)

        - COLOR (color)
        - TON (odor)
        - TURBIDITY (NTU)
- analyte_value
    - mean result of tests performed on this sample for this chemical
- value_qualifier
    - ND means Not Detected or 0 (if applicable)
- value_units
    - units for the value of measurement
        - color
        - mg/L
        - NTU
        - pH (none)
        - ug/L (none)
        - UMHO/CM (none)
        - odor
- source description
    - text description of where sample came from

```
# breaking apart analyte column
Fluoride_mgperL <- ifelse(chem$analyte=="FLUORIDE", chem$analyte_value, 0)
Color_test <- ifelse(chem$analyte=="COLOR", chem$analyte_value, 0)
Ton_odor <- ifelse(chem$analyte=="TON", chem$analyte_value, 0)
Turbidity_NTU <- ifelse(chem$analyte=="TURBIDITY", chem$analyte_value, 0)

chem_nums <- data.frame(Fluoride_mgperL=Fluoride_mgperL,
                        Color_test=Color_test,
                        Ton_odor=Ton_odor,
                        Turbidity_NTU=Turbidity_NTU)
```

```
# histograms
hist.data.frame(chem_nums)
```



```
summary(chem_nums)
```

```
##  Fluoride_mgperL     Color_test        Ton_odor      Turbidity_NTU
##  Min.   :0.00000   Min.   : 0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.00000   Median : 0.000   Median :0.0000   Median :0.0000
##  Mean   :0.06663   Mean   : 0.173   Mean   :0.3246   Mean   :0.0304
##  3rd Qu.:0.00000   3rd Qu.: 0.000   3rd Qu.:1.0000   3rd Qu.:0.0700
##  Max.   :1.07000   Max.   :10.000   Max.   :4.0000   Max.   :3.4000
##  NA's   :21        NA's   :4008     NA's   :34       NA's   :817
```

Drinking Water Chemicals:

Title:

What are you drinking?

Motivation:

Clean drinking water is important and quality of water can change very quickly so timely predictions are vital.

Problem Statement:

Utilize drinking water test results to determine bad drinking water (according to guidelines). Ideally, we want to predict times that the quality suffers so that cause(s) can be found for routine quality issues (if any).

Notable Findings:

We need a better way to split the “analyte” column into separate columns for each measurement so that we don’t have a ton of “zero” values as place holders.