CSI 2300: Intro to Data Science

In-Class Exercise 06: Exploratory Data Analysis

These are the packages that we'll need for today's exercises:

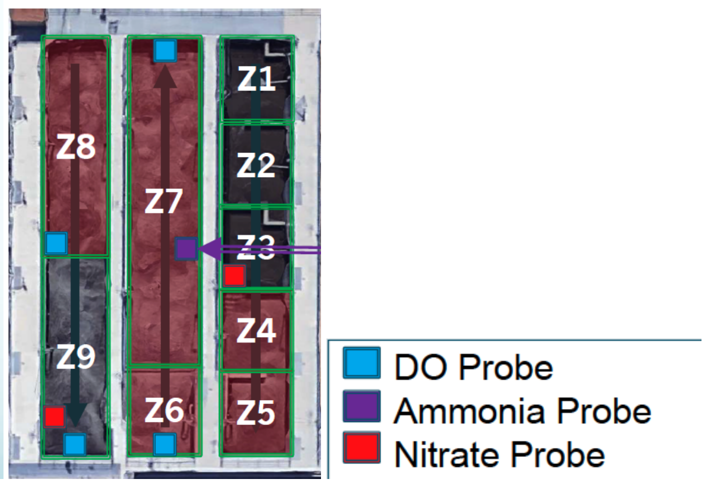
The data for today comes from the mowater package. You can install it by running the code below. You only want this chunk to run once, so after you have the package, set eval back to FALSE.

```r
library(lubridate)
```

The data for today's exercises come from the Boulder Water Resource and Recovery Facility.

First is a picture of the facility where the data are collected. It shows three aeration basins together, and the next plot shows a diagram of the flow of water through one aeration basin. The red highlighted basins are "aerated," meaning that oxygen is being pumped by blowers into the sludge, and the other basins are not aerated. One goal with this data is to try to predict ammonia in Zone 7.

```r
load(file="dat/boulder_ammonia.rda")
```



1. Look at the names of the variables in the data file. Using just the names, can you figure out what each of the variables is? What is the naming convention used?

```r
data <- boulder_ammonia

dim(data) # How many observation I have
# [1] 25908    16

colnames(data)
#  [1] "Date.Time"                 "AB3.Z6.DO.mg.L"
#  [3] "AB3.Z7.DO.mg.L"            "AB3.Z8.DO.mg.L"
#  [5] "AB3.Z9.DO.mg.L"            "AB3.Z6.Header.Flow.SCFM"
#  [7] "AB3.Z7.Header.Flow.SCFM"   "AB3.Z8.Header.Flow.SCFM"
#  [9] "AB3.Zone.6.Valve.Position" "AB3.Zone.7.Valve.Position"
# [11] "AB3.Zone.8.Valve.Position" "AB3.Z7.Ammonia.mg.N.L"
# [13] "AB3.Z3.Nitrate.mg.N.L"     "AB3.Z3.NO2.mg.N.L"
# [15] "AB3.Z9.Nitrate.mg.N.L"     "AB3.Z9.NO2.mg.N.L"

#head(data)
```

2. Describe the type of each of the variables.

```
summary(data)
#    Date.Time                      AB3.Z6.DO.mg.L   AB3.Z7.DO.mg.L
#  Min.   :2019-01-01 00:05:00.00   Min.   :0.0265   Min.   :0.04889
#  1st Qu.:2019-01-23 11:48:45.00   1st Qu.:2.4347   1st Qu.:2.06613
#  Median :2019-02-14 23:32:30.00   Median :2.5512   Median :2.47215
#  Mean   :2019-02-14 23:47:05.53   Mean   :2.6080   Mean   :2.61652
#  3rd Qu.:2019-03-09 11:16:15.00   3rd Qu.:2.7973   3rd Qu.:2.80979
#  Max.   :2019-04-01 00:00:00.00   Max.   :8.4971   Max.   :7.23867
#  AB3.Z8.DO.mg.L   AB3.Z9.DO.mg.L     AB3.Z6.Header.Flow.SCFM
#  Min.   :0.0219   Min.   :0.000000   Min.   :   0
#  1st Qu.:0.9556   1st Qu.:0.000000   1st Qu.:1150
#  Median :1.0094   Median :0.006410   Median :1467
#  Mean   :1.1857   Mean   :0.007371   Mean   :1511
#  3rd Qu.:1.1194   3rd Qu.:0.012500   3rd Qu.:1899
#  Max.   :8.3965   Max.   :1.330875   Max.   :2866
#  AB3.Z7.Header.Flow.SCFM AB3.Z8.Header.Flow.SCFM AB3.Zone.6.Valve.Position
#  Min.   :   4.025        Min.   :  2.4           Min.   :18.32
#  1st Qu.: 422.273        1st Qu.:150.8           1st Qu.:45.41
#  Median : 456.852        Median :272.4           Median :52.68
#  Mean   : 545.264        Mean   :269.4           Mean   :52.25
#  3rd Qu.: 649.732        3rd Qu.:369.5           3rd Qu.:59.47
#  Max.   :1769.409        Max.   :693.0           Max.   :90.00
#  AB3.Zone.7.Valve.Position AB3.Zone.8.Valve.Position AB3.Z7.Ammonia.mg.N.L
#  Min.   : 16.06            Min.   :  4.834           Min.   : 0.1104
#  1st Qu.: 36.08            1st Qu.: 10.773           1st Qu.: 1.4363
#  Median : 39.83            Median : 16.643           Median : 3.4560
#  Mean   : 40.40            Mean   : 16.120           Mean   : 3.9496
#  3rd Qu.: 43.82            3rd Qu.: 20.375           3rd Qu.: 6.0705
#  Max.   :100.20            Max.   :100.083           Max.   :12.4856
#  AB3.Z3.Nitrate.mg.N.L AB3.Z3.NO2.mg.N.L AB3.Z9.Nitrate.mg.N.L
#  Min.   : 0.00000      Min.   : 0.6438   Min.   : 0.6438
#  1st Qu.: 0.02118      1st Qu.: 6.2694   1st Qu.: 6.2694
#  Median : 0.96369      Median : 7.5796   Median : 7.5796
#  Mean   : 2.75315      Mean   : 7.7932   Mean   : 7.7932
#  3rd Qu.: 3.57012      3rd Qu.: 9.2462   3rd Qu.: 9.2462
#  Max.   :43.99928      Max.   :17.0555   Max.   :17.0555
#  AB3.Z9.NO2.mg.N.L
#  Min.   :0.000000
#  1st Qu.:0.006050
#  Median :0.008775
#  Mean   :0.136584
#  3rd Qu.:0.167637
#  Max.   :2.522425

#DATA IS QUANTITAVITE AND
#CONTIUOUS
```

3. How frequently are the measurements taken?

4. What are the first and last dates in the dataset?

```r
head(data$Date.Time) #2019-01-01
# [1] "2019-01-01 00:05:00 UTC" "2019-01-01 00:10:00 UTC"
# [3] "2019-01-01 00:15:00 UTC" "2019-01-01 00:20:00 UTC"
# [5] "2019-01-01 00:25:00 UTC" "2019-01-01 00:30:00 UTC"
tail(data$Date.Time) #2019-03-31
# [1] "2019-03-31 23:35:00 UTC" "2019-03-31 23:40:00 UTC"
# [3] "2019-03-31 23:45:00 UTC" "2019-03-31 23:50:00 UTC"
# [5] "2019-03-31 23:55:00 UTC" "2019-04-01 00:00:00 UTC"

min(data$Date.Time) #2019-01-01
# [1] "2019-01-01 00:05:00 UTC"
max(data$Date.Time) #2019-04-01
# [1] "2019-04-01 UTC"
```

5. Compute the mean, median, and standard deviation of the ammonia data. To identify observations that are unusual, people commonly compute the number of standard deviations away from the mean that an observation is. Find the minimum and maximum values of ammonias, and compute the number of standard deviations these values are away from the mean.

```r
ammonia <- data$AB3.Z7.Ammonia.mg.N.L

mean(ammonia) #3.949557
# [1] 3.949557
median(ammonia) #3.45602
# [1] 3.45602
sd(ammonia) #2.752019
# [1] 2.752019

min(ammonia) #0.1103667
# [1] 0.1103667
max(ammonia) #12.48557
# [1] 12.48557

min(ammonia) - max(ammonia)/sd(ammonia) #-4.426508
# [1] -4.426508

(max(ammonia) - mean(ammonia))/sd(ammonia) #3.101726
# [1] 3.101726
```

6. Compute the 1, 5, 10, and 90, 95, 99$^{\text{th}}$ quantiles of ammonia.

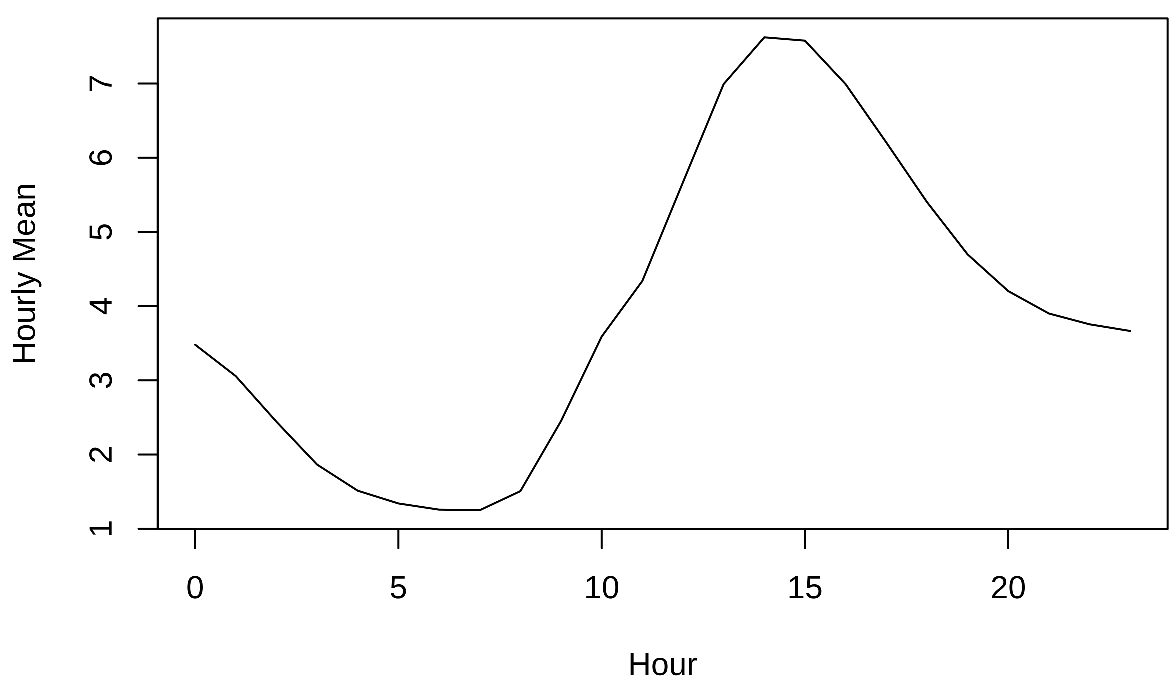
```r
quantile(ammonia, c(0.01, 0.05, 0.1, 0.9, 0.95, 0.99))
#        1%        5%       10%       90%       95%       99%
#  0.2933933  0.5134684  0.7636366  7.9711632  8.9596873 10.4291826
```

1%	5%	10%	90%	95%	99%
0.2933933	0.5134684	0.7636366	7.9711632	8.9596873	10.4291826

7. The way to obtain the hour associated with each observation is given below. Note that the hours are labeled as {0,1,2,. . . 23}. Find the mean value of ammonia for each hour of the day. The command

tapply() could be useful here. Do there appear to be differences in ammonia across the course of a day? If so, why do you think that this could be occurring?

```
hour <- hour(boulder_ammonia$Date.Time)

hour0 <- ammonia[which(hour==0)]

round(mean(hour0), 2) #mean at hour0 is: 3.48
# [1] 3.48


#FOR LOOP:

mean_by_hour <- NULL

for (i in 0:23){
  mean_by_hour[i+1] <- round(mean(ammonia[which(hour==i)]), 2)
}

mean_by_hour <- tapply(ammonia, hour, mean)

plot(0:23, mean_by_hour, type = "l", xlab = "Hour", ylab = "Hourly Mean")
```



8. Compute the mean and standard deviation of the dissolved oxygen (DO) as you move from Zone 6 to Zone 9. How does DO change as you move through the basin?

```r
mean(data$AB3.Z6.DO.mg.L)
# [1] 2.607962
mean(data$AB3.Z7.DO.mg.L)
# [1] 2.616522
mean(data$AB3.Z8.DO.mg.L)
# [1] 1.185695
mean(data$AB3.Z9.DO.mg.L)
# [1] 0.007370941
```

As you move through the basin from Zone 6 to Zone 9, the dissolved oxygen levels decrease significantly, with the most pronounced drop occurring between Zone 7 and Zone 8, and reaching near-zero levels in Zone 9. This pattern indicates that the basin becomes increasingly oxygen-depleted as you move from the earlier zones (6 and 7) to the later zones (8 and 9).

9. Both the mean and the median measure the center of a dataset. However, there can be differences between them. If a distribution is symmetric around its center, the mean and the median will be about the same. If the distribution is not symmetric, the mean will be drawn to the more extreme values. Compare the mean and median of nitrate in both Zone 3 versus Zone 9. In which zone does the distribution of nitrate appear to be symmetric, based only on comparing their mean and median?

```r
mean(data$AB3.Z3.Nitrate.mg.N.L)
# [1] 2.753146
median(data$AB3.Z3.Nitrate.mg.N.L)
# [1] 0.9636916

mean(data$AB3.Z9.Nitrate.mg.N.L)
# [1] 7.793221
median(data$AB3.Z9.Nitrate.mg.N.L)
# [1] 7.579566
```

Zone 3: The distribution of nitrate is not symmetric, as the mean is much higher than the median.

Zone 9: The distribution of nitrate is symmetric, as the mean and median are very close.

10. How often is ammonia above 8 mg/L in this dataset?

```r
length(which(ammonia > 8))
# [1] 2547
length(ammonia)
# [1] 25908
```

Ammonia levels go above 8 mg/L in about 9.83% of the data. This means it happens less than 10% of the time, so it's not very common.