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:

```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("~/Documents/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
colnames(boulder_ammonia)
#  [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"
```

The Date.time variable likely shows the date/time that the rest of the variables were recorded. AB stands for "aeration basin" while z represents the zone. The DO represents the "dissvolved oxygen." The naming convention that is utilized is snake case. boulder_ammonia$

2. Describe the type of each of the variables.

The Date.time variable is an equally-spaced time index. The rest are numeric.

3. How frequently are the measurements taken?

The measurements are taken every 5 minutes.

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

```r
head(boulder_ammonia, 1)
#             Date.Time AB3.Z6.DO.mg.L AB3.Z7.DO.mg.L AB3.Z8.DO.mg.L
# 1 2019-01-01 00:05:00       2.505239       2.332144        0.88655
#   AB3.Z9.DO.mg.L AB3.Z6.Header.Flow.SCFM AB3.Z7.Header.Flow.SCFM
# 1       0.014775                2074.175                618.2874
#   AB3.Z8.Header.Flow.SCFM AB3.Zone.6.Valve.Position AB3.Zone.7.Valve.Position
# 1                 223.743                  58.04692                  33.01925
#   AB3.Zone.8.Valve.Position AB3.Z7.Ammonia.mg.N.L AB3.Z3.Nitrate.mg.N.L
```

```
# 1                  12.84363             0.7907111              2.317133
#   AB3.Z3.NO2.mg.N.L AB3.Z9.Nitrate.mg.N.L AB3.Z9.NO2.mg.N.L
# 1          9.355388              9.355388           0.01015
tail(boulder_ammonia, 1)
#        Date.Time AB3.Z6.DO.mg.L AB3.Z7.DO.mg.L AB3.Z8.DO.mg.L AB3.Z9.DO.mg.L
# 25908 2019-04-01       2.646196       4.049265       1.981365              0
#       AB3.Z6.Header.Flow.SCFM AB3.Z7.Header.Flow.SCFM AB3.Z8.Header.Flow.SCFM
# 25908                1506.474                811.8342                95.86256
#       AB3.Zone.6.Valve.Position AB3.Zone.7.Valve.Position
# 25908                  46.51333                  67.08715
#       AB3.Zone.8.Valve.Position AB3.Z7.Ammonia.mg.N.L AB3.Z3.Nitrate.mg.N.L
# 25908                  9.739666              1.238144               5.26771
#       AB3.Z3.NO2.mg.N.L AB3.Z9.Nitrate.mg.N.L AB3.Z9.NO2.mg.N.L
# 25908          10.82222              10.82222          0.130625
```

First date is 2019-01-01, the last date is 2019-04-01.

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.

```
ammonia <- boulder_ammonia$AB3.Z7.Ammonia.mg.N.L
mean(ammonia)
# [1] 3.949557
median(ammonia)
# [1] 3.45602
sd(ammonia)
# [1] 2.752019

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

(min(ammonia) - mean(ammonia))/sd(ammonia) # Minimum is 1.4 sd below the mean
# [1] -1.395045
(max(ammonia) - mean(ammonia))/sd(ammonia) # Maximum is 3.1 sd above the mean
# [1] 3.101726
```

6. Compute the 1, 5, 10, and 90, 95, 99th quantiles of ammonia.

```
quantile(ammonia) # This returns the quartiles
#         0%        25%        50%        75%       100%
#  0.1103667  1.4362884  3.4560203  6.0704888 12.4855664
quantile(ammonia, probs = c(01,.05,.1,.9,.99))
#       100%         5%        10%        90%        99%
# 12.4855664  0.5134684  0.7636366  7.9711632 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?

```r
hour <- hour(boulder_ammonia$Date.Time)
mean_ammonia_hour <- tapply(ammonia, hour, mean)
```

During the beginning of the day, the mean decreases. After it gets to the seventh hour, the mean starts rising until after the 15th hour, it starts decreasing. This could be occurring because people use more water in the mornings and the middle of the day compared to later at night.

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

sd(boulder_ammonia$AB3.Z6.DO.mg.L)
# [1] 0.2853541
sd(boulder_ammonia$AB3.Z7.DO.mg.L)
# [1] 0.7296249
sd(boulder_ammonia$AB3.Z8.DO.mg.L)
# [1] 0.5957249
sd(boulder_ammonia$AB3.Z9.DO.mg.L)
# [1] 0.01732953
```

The mean of the DO generally decreases as you move from zone 6 to 9. The standard deviation decreases generally between the different zones but peaks in zone 7.

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(boulder_ammonia$AB3.Z3.Nitrate.mg.N.L)
# [1] 2.753146
median(boulder_ammonia$AB3.Z3.Nitrate.mg.N.L)
# [1] 0.9636916

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

The zone where nitrate appears more symmetric is in zone 9.

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

```r
length(which(ammonia>8))
# [1] 2547
(length(which(ammonia>8))/length(ammonia)) * 100
# [1] 9.83094
```

Ammonia is above 8 mg/L in this dataset 2547 times, or 9.83% of the time