CSI 2300: Intro to Data Science

In-Class Exercise 09: Exploratory Data Analysis

The data for today's exercises come from the `mowater` library, the `eml` dataset. This data are about measurements of the properties of water in the Eagle Mountain Lake reservoir in North Texas.

1. Load the `mowater` library, and then load the `eml` dataset. In the RMarkdown document, show the commands you use to do this, but not the output of those commands (`message=FALSE` as an option in the code chunk header is a good way to do this).

```
library(devtools)
#devtools::install_github("rachwhatsit/mowater_pkg", subdir = "mowateR")
library(mowateR)
data(eml)
#help(eml)

load("dat/eml.rda")
head(eml)
#    Date.Time Depth   Temp     DO   DOsat    pH    Cond
# 1 2019-04-25   0.0 19.156 10.455 116.370 8.586 421.801
# 2 2019-04-25   0.5 19.137 10.468 115.732 8.578 421.859
# 3 2019-04-25   1.0 19.193 10.411 115.345 8.617 419.710
# 4 2019-04-25   1.5 19.229 10.414 115.121 8.618 419.609
# 5 2019-04-25   2.0 19.171 10.419 114.568 8.571 421.432
# 6 2019-04-25   2.5 19.239 10.351 114.830 8.570 421.246
```

2. Next, inspect the dataset.

```
#dim(eml)
nrow(eml)#number of observations
# [1] 35532
ncol(eml)# 7 variables
# [1] 7
sum(sort(eml$Date.Time) == eml$Date.Time) #check if sorted
# [1] 35532

summary(eml)
#    Date.Time                       Depth           Temp             DO
#  Min.   :2019-04-25 00:00:00   Min.   : 0.0   Min.   :17.72   Min.   : 0.000
#  1st Qu.:2019-05-30 05:30:00   1st Qu.: 2.5   1st Qu.:24.31   1st Qu.: 2.070
#  Median :2019-07-04 11:00:00   Median : 5.0   Median :27.40   Median : 4.958
```

```
#  Mean   :2019-07-04 11:00:00   Mean   : 5.0   Mean   :26.24   Mean   : 4.795
#  3rd Qu.:2019-08-08 16:30:00   3rd Qu.: 7.5   3rd Qu.:28.79   3rd Qu.: 7.255
#  Max.   :2019-09-12 22:00:00   Max.   :10.0   Max.   :34.15   Max.   :15.508
#      DOsat               pH              Cond
#  Min.   :  0.00   Min.   :6.938   Min.   :290.4
#  1st Qu.: 26.22   1st Qu.:7.599   1st Qu.:347.1
#  Median : 62.04   Median :7.998   Median :360.4
#  Mean   : 59.82   Mean   :7.983   Mean   :362.8
#  3rd Qu.: 88.37   3rd Qu.:8.298   3rd Qu.:368.3
#  Max.   :218.65   Max.   :9.551   Max.   :439.8
min(eml$Depth)
# [1] 0
max(eml$Depth)
# [1] 10

range(eml$Depth)
# [1]  0 10

sum(is.na(eml)) # 0, no missing values
# [1] 0
```

- How many variables are there?

7

- How many observations are there?

35532

- Are the data already sorted in time order?

yes, they are in time order.

```
sum(sort(eml$Date.Time) == eml$Date.Time)#yes, they are in time order
# [1] 35532
```

- What do the variables represent, do you think? You may want to consult the `help` for the dataset to understand it better, including the units of measurement.

Date time is date and two-hour time period, Depth is profile measured in meters TODO: copy over for the variable type - What are the ranges of the values?

```
#summary(eml) #one way
#range(eml$Depth) #another way, if repeated for eachcol
apply(eml,2, range)
#      Date.Time             Depth  Temp      DO        DOsat      pH
# [1,] "2019-04-25 00:00:00" " 0.0" "17.7160" " 0.0000" "  0.0000" "6.9380"
# [2,] "2019-09-12 22:00:00" "10.0" "34.1510" "15.5080" "218.6450" "9.5510"
#      Cond
# [1,] "290.4400"
# [2,] "439.8370"
```

- Are there any missing values (NA values)?

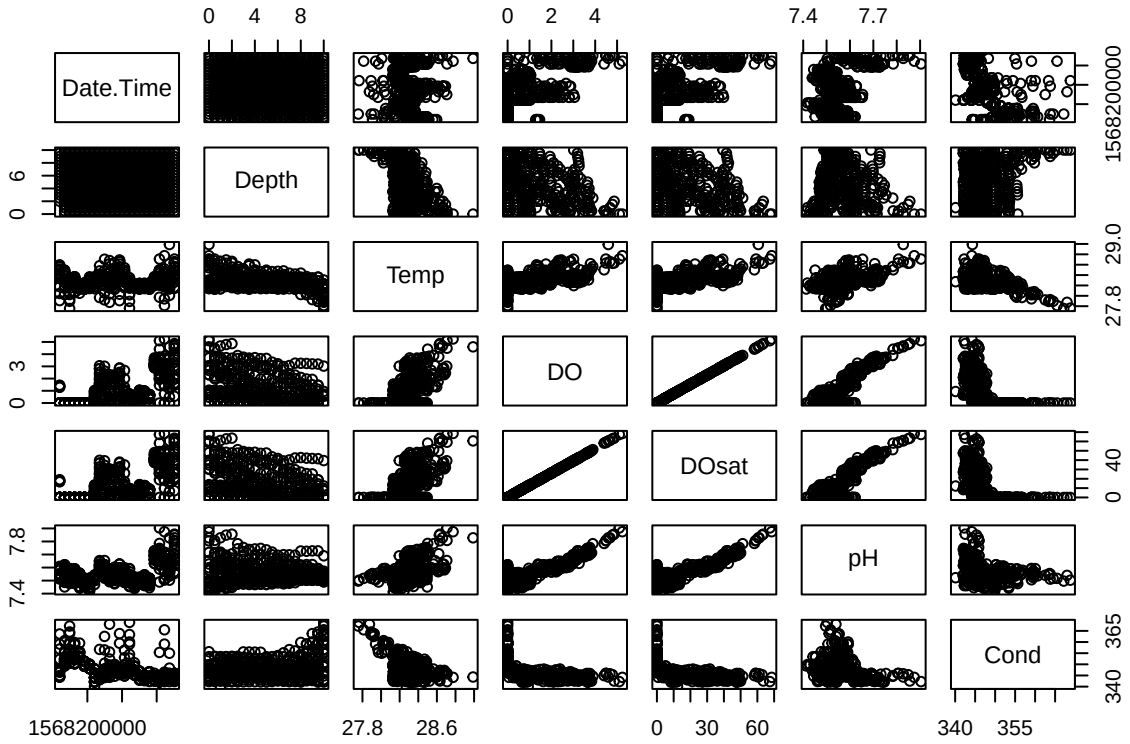
```
sum(is.na(eml))#no missing values
# [1] 0
```

3. The size of our dataset is rather large for easy (and fast) manipulation. Pare it down by creating a new data frame with only the last 500 observations in it. This is throwing away a lot of information, and we should be careful any time we do this. But if we try to use all the data, it may be too time-consuming for an in-class exercise.

```
eml_small <- tail(eml, 500)
```

4. Now that the data are of a manageable size for visualizing it, plot all of the variables against each other. Try calling `plot(eml_small)`, but replace `eml_small` with the name of the *small* data frame you just created. This creates a matrix of pairwise scatterplots.
   (If this takes a very long time, you may have made the mistake of trying to plot the original dataset, which is quite large for this task.)

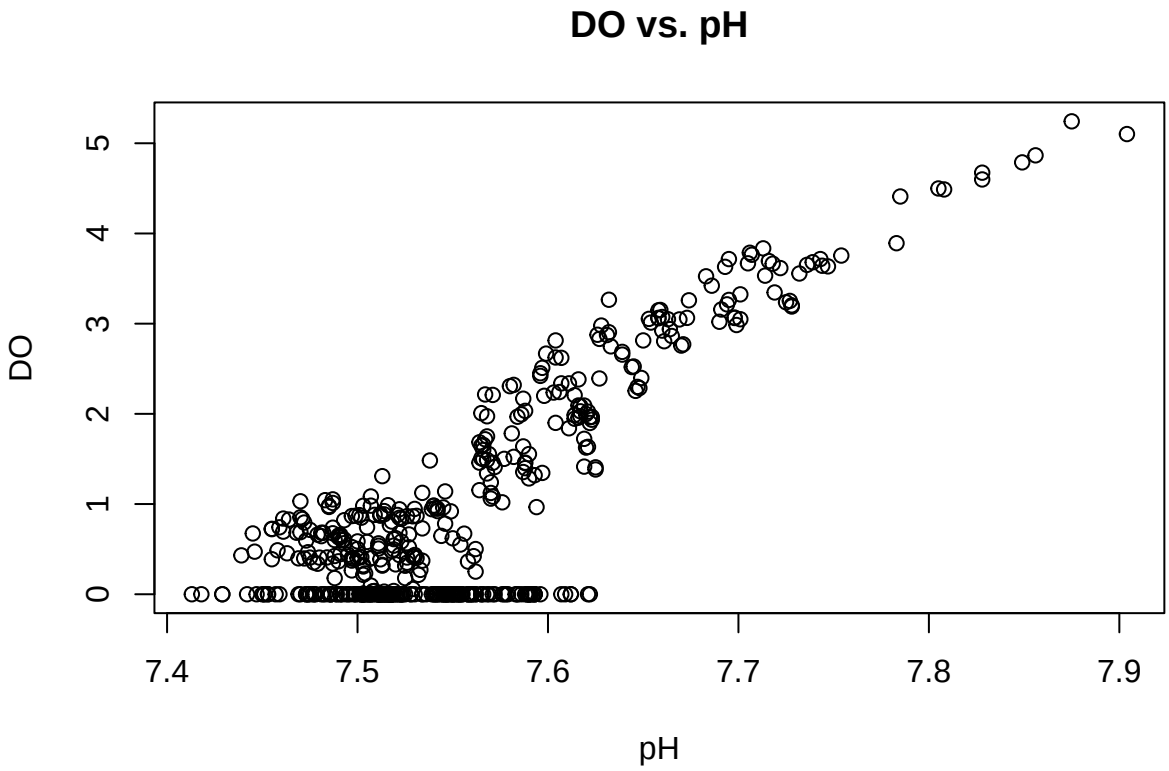
```
plot(eml_small)
```



5. Look at the plot and consider which pairs of variables appear to be linearly related. Choose `DO` and `pH`. Make a scatter plot of `DO` on the y-axis and `pH` on the x-axis. Which one are you thinking of as the independent variable, and which one are you thinking of as the dependent variable?

TODO: list linear relationship

```
plot(eml_small$pH, eml_small$DO,
     xlab = "pH",
     ylab = "DO",
     main = "DO vs. pH")
```



TODO: tell me which is the dependent and independent variable

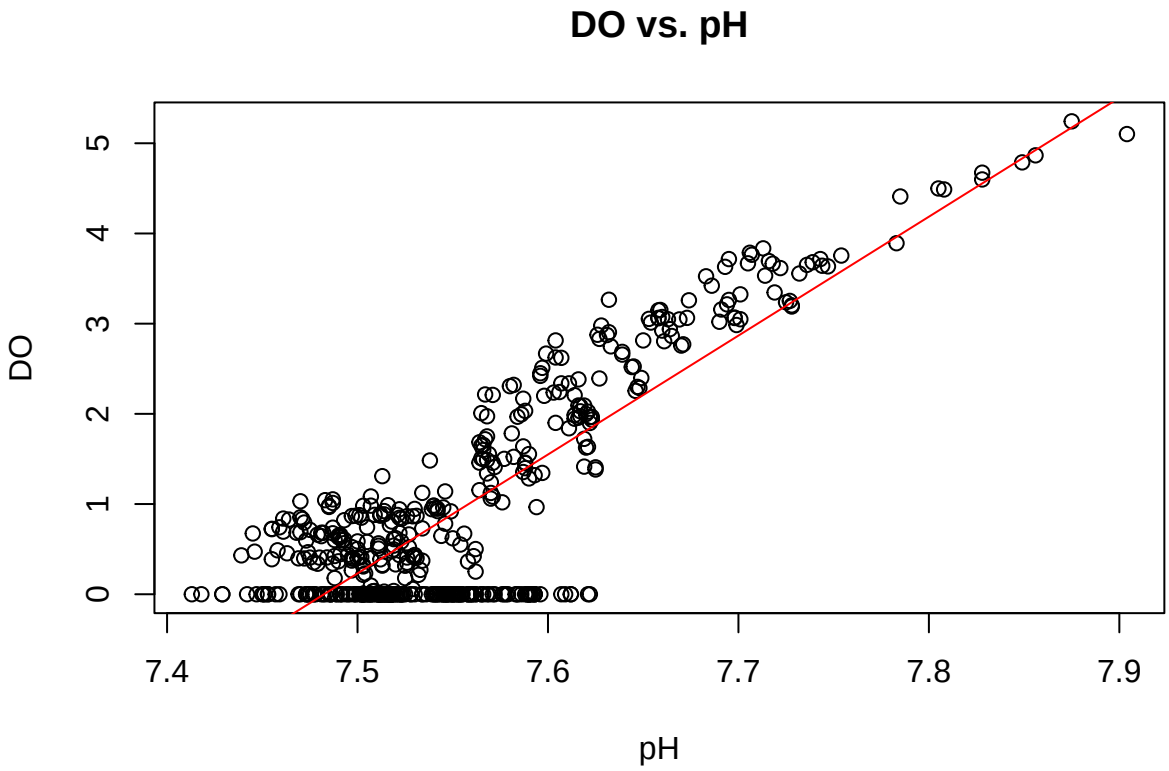
pH is the independent variable DO is the dependent variable

6. Run a linear regression by calling `lm(DO ~ pH, data=eml_small)` where again `eml_small` is the small data frame. Plot the regression line on the scatter plot using `abline` on the model that `lm` returns. If the line doesn't appear to follow the data, you may have switched the variables (for the plot versus the linear regression).

```
m <- lm(DO ~ pH, data=eml_small)

# Plot the scatter plot again for clarity
plot(eml_small$pH, eml_small$DO,
     xlab = "pH",
     ylab = "DO",
     main = "DO vs. pH")

abline(m, col = 'red')
```



7. Investigate the coefficients and summary statistics of the model that `lm` gave you. Comment on the coefficient values, significance levels of the coefficients (are they significantly different from 0), and R^2 values. Does there appear to be a linear relationship between these two variables?

```
#sign and magnitude of each coef in terms of the values modeled, e.g.
#as pH increases by one standard pH unit, DO increases by about 13 mg/L
#intercept is -98, if we imagine pH = 0, then DO = -98 mg/L

# Display the coefficients of the model
coef(m)
# (Intercept)          pH
#   -98.63370    13.18199

# Provide a detailed summary of the model including significance and R-squared
summary(m)
#
# Call:
# lm(formula = DO ~ pH, data = eml_small)
#
# Residuals:
```

```
#      Min       1Q   Median       3Q      Max
# -1.83942 -0.41907  0.09877  0.48027  1.29376
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept) -98.6337     2.8287  -34.87   <2e-16 ***
# pH           13.1820     0.3743   35.22   <2e-16 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 0.6543 on 498 degrees of freedom
# Multiple R-squared:  0.7135,  Adjusted R-squared:  0.713
# F-statistic:  1240 on 1 and 498 DF,  p-value: < 2.2e-16
```

The model shows a strong linear relationship between pH and DO, with pH having a significant positive effect on DO. As the intercept is -98, even though it should be >0, with the date range it shows the intercept and pH coefficient are highly significant. It also shows that there is a positive correlation so when pH goes up DO goes up and vice versa. With 0.0001 alpha level it shows the coef is significant.

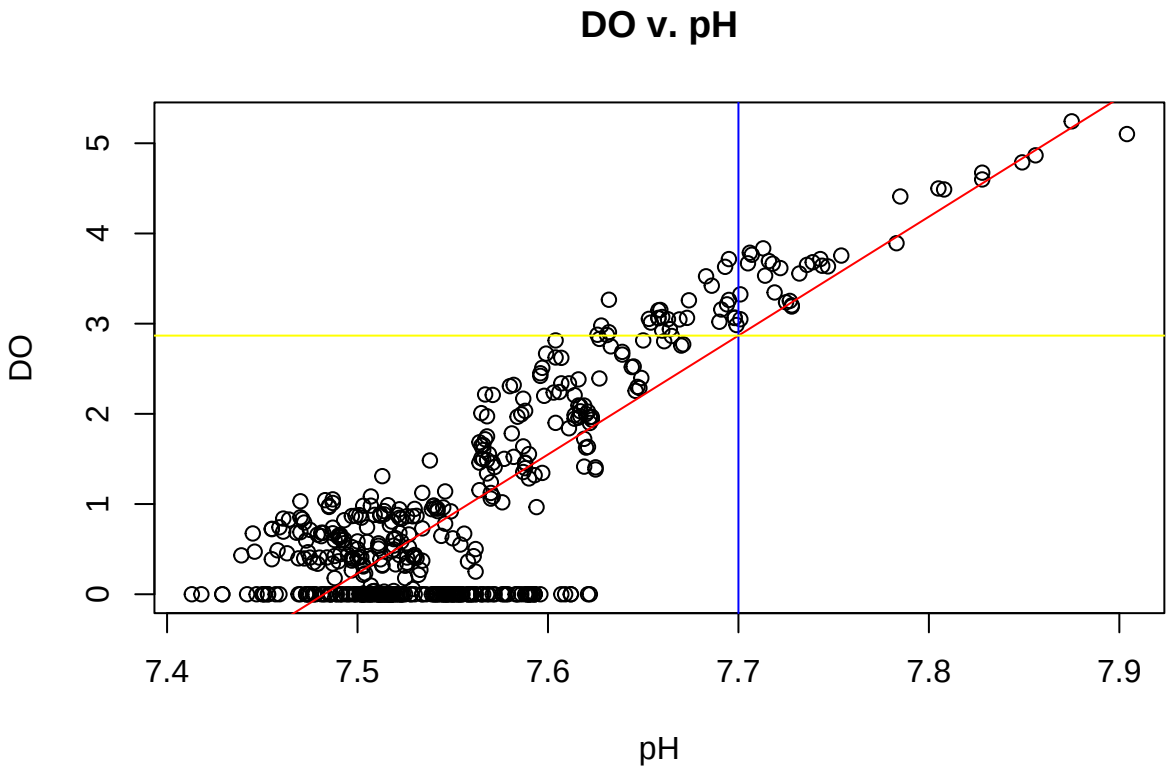
8. Use the model to make a prediction of the dependent variable for when the variable pH is 7.7. You can do this by direct computation if you want (rather than using any specialty R command). Hand-check your work on your plot. Is the predicted value close to the plotted data?

```
pH_value <- 7.7

y_new <- coef(m)[1] + coef(m)[2] * pH_value

plot(eml_small$pH, eml_small$DO,
     xlab = "pH",
     ylab = "DO",
     main = "DO v. pH")
abline(m, col = "red")

abline(v = pH_value, col = "blue")  # pH = 7.7 line
abline(h = y_new, col = "yellow")      # Predicted DO line
```



9. Do you trust this model to make a prediction of `DO` for a `pH` value of 3? Explain your answer.

No as our model has only been trained on values between pH 7.4 to 7.9 so the model is not reliable for values outside of this range.

10. Extra credit: repeat steps 5-7, but for a different pair of variables than the pair you were just working with. Compare the models; does it appear that one pair of variables is more strongly linearly related than the other pair? Note, do not choose the pair `DO` and `DOsat` as they are the same variable but measured in different units.